CREATING A VERSATILE TOOLKIT FOR TRANSGENE EXPRESSION USING BACS AND FOR DISSECTING LARGE-SCALE CHROMATIN ORGANIZATION

BY

BINHUI ZHAO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Cell and Developmental Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Andrew S. Belmont, Chair and Director of Research
Professor Lisa J. Stubbs
Professor Brian C. Freeman
Professor Huimin Zhao
Associate Professor Supriya Prasanth

ABSTRACT

In eukaryotes, the genetic material, DNA, is highly compacted with histone proteins to form chromatin in interphase nuclei. Both the higher levels of chromatin folding and the spatial organization of the chromatin, referred to as large-scale chromatin organization, have been shown to correlate with transcriptional activity. One example suggesting transcriptional regulation by large-scale chromatin organization is position effects and position effect variegations observed in transgene expressions, which, as well as other epigenetic silencing mechanisms, have been obstacles to achieving predictable and stable transgene expression. Molecular dissections of the determinants regulating large-scale chromatin organization would help to elucidate the real relationship between large-scale chromatin organization and transcriptional regulation, yet are difficult due to the complexity of the mammalian genome. Bacterial artificial chromosomes (BACs), containing 100-300 kb mammalian genomic regions have been shown to recapitulate the expression level and nuclear localization of the corresponding genomic regions, and to protect embedded reporter mini-genes from epigenetic silencing.

Here in Chapter 2 we show that BACs could provide a versatile platform for achieving reproducible, stable simultaneous expression of multiple transgenes maintained either as episomes or stably integrated copies. Moreover, in Chapter 3 we show that BACs could be used as a powerful model system for dissecting mechanisms regulating large-scale chromatin organization, by demonstrating distinctive large-scale chromatin conformations formed by BAC transgene arrays and results indicating separation of large-scale chromatin compaction, nuclear localization and transcriptional activities.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

TRANSGENE EXPRESSION

Transgene expression is a fundamental technique in both basic biological research and biotechnology. Yet most of the widely used transgene expression systems, such plasmids, lentiviruses and transposons , still suffer from various degrees of transgene silencing. The integration site of the transgene has a great effect on its expression, known as chromosome position effects (Akhtar *et al.*, 2013; Chen *et al.*, 2013). Transgenes integrated into inactive chromatin regions tend to express at low levels or have variegated expression (position effect variegation, PEV) (Robertson *et al.*, 1995; Ramunas *et al.*, 2007; Girton and Johansen, 2008). However, integration into active chromatin regions is not sufficient for high level of expression, as prokaryotic and viral sequences, as well as transgene concatemers can become targets for epigenetic silencing (Dorer and Henikoff, 1997; Hong, Sherley and Lauffenburger, 2001; Chen *et al.*, 2004) Thus, it usually takes tedious screening to get clones with desired levels of transgene expression. However, transgene expression could deteriorate with increasing number of passages or freezes, or when the cell clones get differentiated.

A conventional way to protect the transgene from silencing is to include certain cis-elements in the expression cassette. Various cis-elements have been shown to alleviate transgene silencing. Examples include insulators (Pikaart, Recillas-Targa and Felsenfeld, 1998; Emery *et al.*, 2000), locus control regions (LCRs) (Grosveld *et al.*, 1987; Guy *et al.*, 1996), scaffold/matrix attachment regions (S/MARs) (Phi-Van *et al.*,

1990; Kim *et al.*, 2004), ubiquitous chromatin opening elements (UCOEs) (Williams *et al.*, 2005; Müller-Kuller *et al.*, 2015) and anti-repressors (Kwaks *et al.*, 2003).

Recently, bacterial artificial chromosomes (BACs) have been shown to be more effective than the conventional cis-elements for achieving sustained high-level transgene expressions (Blaas *et al.*, 2009; Bian and Belmont, 2010). BACs usually contain 100-300 kb mammalian genomic regions and can integrate into mammalian genomes as tandem repeats. It is hypothesized that the large BAC transgene array could shield the embedded mini-genes from influences from the chromosomal integration sites and that the mini-gene expression is determined by the properties of the genomic region contained within the BACs. Based on this hypothesis, a study explored different promoter and BAC combinations for high level production of recombinant proteins (Zboray *et al.*, 2015).

EPISOMAL TRANSGENE EXPRESSION SYSTEMS

In some transgene expression applications, it is desirable for the transgenes to persist in the host cell nuclei as extrachromosomal DNA, instead of integrating into the host chromosomes, to avoid possible disruption of host genes, and/or to eventually eliminate these transgenes from the cells. Various episomal vectors have been developed (Van Craenenbroeck, Vanhoenacker and Haegeman, 2000; Conese, Auriche and Ascenzioni, 2004; Ehrhardt *et al.*, 2008; Lufino, Edser and Wade-Martins, 2008). Viral sequence based episomal vectors rely on viral proteins and a viral replication origin for replication and partitioning into daughter cells (Milanesi *et al.*, 1984; Rawlins *et al.*, 1985; Yates, Warren and Sugden, 1985; Piirsoo *et al.*, 1996). Epstein Barr virus (EBV)

derived episomal vectors are lost at a low rate without selection (Nanbo, Sugden and Sugden, 2007) and have been used to produce transgene-free iPS cells (Yu *et al.*, 2009). The main disadvantage of viral based vectors is the requirement for the viral proteins, which could lead to transformation of the transfected cells (Frappier, 2012). None viral episomal vectors have two key components: a S/MAR sequence from the human β-interferon gene cluster, and a upstream transcription going through the S/MAR (Piechaczek *et al.*, 1999; Baiker *et al.*, 2000; Jenke *et al.*, 2004). S/MAR based episomal plasmids are maintained at low copy number and are mitotically stable without selection (Piechaczek *et al.*, 1999; Baiker *et al.*, 2000; Jenke *et al.*, 2004; Stehle *et al.*, 2007; Argyros *et al.*, 2008).

Human artificial chromosomes (HACs) are a high capacity episomal vector system. They are usually 1-10 Mb in size, and contain centromeric repeat sequences for episomal status maintenance (Harrington *et al.*, 1997; Mills *et al.*, 1999; Kazuki and Oshimura, 2011; Kouprina *et al.*, 2013). Most HACs are constructed by either "top down" or "bottom up" approaches. In the "top down" approach, a natural chromosome is fragmented by a targeting vector containing a terminal telomere segment, a selectable marker, and sometimes a region of homology to the target chromosome, generating an engineered mini-chromosome. In the "bottom up" approach, cloned centromeric DNA and genomic DNA, or engineered BACs, PACs, or YACs containing centromeric DNA are introduced into cells to form de novo chromosomes. Telomeric DNA is not necessary, as the de novo chromosomes formed are usually circular DNA (Mejía *et al.*, 2001; Ebersole, 2002; Conese, Auriche and Ascenzioni, 2004). Both the "top down" and the "bottom up" approaches require specific cell lines, and the transfer of the HACs from

the donor cells into the recipient cells is even more difficult (Fournier and Ruddle, 1977; Liskovykh *et al.*, 2016). HACs have shown great potential in a wide range of applications, such as recombinant protein production, drug selection and gene therapy (Kazuki *et al.*, 2010; Takahashi *et al.*, 2010; Hiratsuka *et al.*, 2011; Kim *et al.*, 2011).

DOMAIN ORGANIZATION OF THE INTERPHASE CHROMATIN

The position effects and position effect variegations suggest regulation of transcription at the chromatin level. An interesting hypothesis is that besides transcription factors, an additional level of gene expression regulation exists that acts over 100s-1000s kb sized chromosomal domains. A study reporter showed constructs integrated at 90 different chromosomal positions obtain expression levels that correspond to the activity of the domains of integration (Gierman *et al.*, 2007). Such domains were identified by testing various window sizes to find the highest correlation between reporter gene expression and domain transcriptional activity (Gierman *et al.*, 2007). More studies are needed to verify the hypothesis.

Both microscopy and genomic mapping studies have suggested domain organization of the interphase chromatin. Transmission electron microscopy (TEM) studies with DNA specific stains (Belmont *et al.*, 1989; Bohrmann and Kellenberger, 1994; Bazett-Jones and Hendzel, 1999), as well as recently developed ChromEMT (Ou *et al.*, 2017), have shown most nuclear DNA is present in large-scale, frequently fiber-like structures consisting of ~Mbp size segments with a range of diameters. Moreover, DNA FISH and light microscopy based experiments have shown that each chromosome

occupies a distinct territory, termed chromosome territories (Cremer and Cremer, 2001; Parada *et al.*, 2002).

Such domain organization is consistent with recent discoveries from genome-wide mapping studies. For example, the A/T content of the mammalian genomes form long DNA stretches (>100 kb) termed isochores, with A/T-poor regions roughly corresponding to high gene density regions along the genome (Caron *et al.*, 2001; Costantini, 2006).

Enrichment of certain histone modifications also form 10s-100s kb of domains. H3K27me3 modifications have been shown to form 10s of kb BLOCs (broad local enrichments) over silent genes and intergenic regions in mouse embryonic fibroblast (MEF) cell lines (Pauler *et al.*, 2009). H3K9me2 has also been found to form 10s kb sized domains, along the genome in mouse cells and tissues, which are termed large organized chromatin K9 modifications (LOCKs), and together cover up to 45% of the genome, depending on the cell type (Wen *et al.*, 2009). Analyzing a set of histone modifications over 100s kb-Mb sized intervals have categorized five chromatin states (Zhu *et al.*, 2013).

Moreover, genomic mapping studies have also identified 100 kb-10 Mb nuclear lamina associating domains (LADs), which are generally low in transcriptional activity, and are enriched with repressive histone modifications (Guelen *et al.*, 2008; Peric-Hupkes *et al.*, 2010).

Recent development of genome-wide chromosome conformation capture techniques (3C, 4C, 5C, and Hi-C) has shown that chromatin interactions are spatially restricted into repeated chromatin domains (Dixon *et al.*, 2012; Dixon, Gorkin and Ren,

2016), termed topologically associated domains (TADs). TADs have been found to associated with distinct patterns of histone marks and can be segregate into six subcompartments (Rao *et al.*, 2014).

In spite of the chromatin domains discovered by both TEM and genomic mapping studies, how are these domains established and how do they regulate each other has not been established. A study (Boettiger *et al.*, 2016) showed that 100s kb sized genomic domains with active histone modifications have more "open" chromatin conformation than genomic regions with repressive histone modifications, and genomic regions with neither active or repressive histone modifications have chromatin conformation in between with DNA FISH and super-resolution light microscopy, indicating possible regulation of large-scale chromatin compaction by epigenetic modifications. Moreover, using a multiplexed FISH method to sequentially image many genomic regions a study traced several TADs in human cells and found that TADs are largely organized into two compartments spatially arranged in a polarized manner in individual chromosomes (Wang *et al.*, 2016). Interestingly, a 4.3-Mb region on mouse chromosome 14 that contains four clusters of genes separated by gene "deserts" was shown to form nonrandom "higher order" structures, where the gene clusters and gene deserts regions were always separated from each other (Shopland *et al.*, 2006).

REFERENCES

Akhtar, W. *et al.* (2013) 'Chromatin position effects assayed by thousands of reporters integrated in parallel.', *Cell*. Elsevier, 154(4), pp. 914–27.

Argyros, O. *et al.* (2008) 'Persistent episomal transgene expression in liver following delivery of a scaffold/matrix attachment region containing non-viral vector.', *Gene therapy*, 15(24), pp. 1593–1605.

Baiker, A. *et al.* (2000) 'Mitotic stability of an episomal vector containing a human scaffold/matrix-attached region is provided by association with nuclear matrix.', *Nature cell biology*, 2(3), pp. 182–4.

Bazett-Jones, D. P. and Hendzel, M. J. (1999) 'Electron spectroscopic imaging of chromatin.', *Methods (San Diego, Calif.)*, 17(2), pp. 188–200.

Belmont, A. S. *et al.* (1989) 'Large-scale chromatin structural domains within mitotic and interphase chromosomes in vivo and in vitro.', *Chromosoma*, 98(2), pp. 129–43.

Bian, Q. and Belmont, A. S. (2010) 'BAC TG-EMBED: one-step method for high-level, copy-number-dependent, position-independent transgene expression.', *Nucleic acids research*, 38(11), p. e127.

Blaas, L. *et al.* (2009) 'Bacterial artificial chromosomes improve recombinant protein production in mammalian cells.', *BMC biotechnology*, 9, p. 3.

Boettiger, A. N. *et al.* (2016) 'Super-resolution imaging reveals distinct chromatin folding for different epigenetic states.', *Nature*. Nature Publishing Group, 529(7586), pp. 418–22.

Bohrmann, B. and Kellenberger, E. (1994) 'Immunostaining of DNA in electron microscopy: an amplification and staining procedure for thin sections as alternative to gold labeling.', *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, 42(5), pp. 635–43.

Caron, H. *et al.* (2001) 'The human transcriptome map: clustering of highly expressed genes in chromosomal domains.', *Science (New York, N.Y.)*, 291(5507), pp. 1289–92.

Chen, M. *et al.* (2013) 'Decoupling Epigenetic and Genetic Effects through Systematic Analysis of Gene Position.', *Cell reports*. Elsevier, 3(1), pp. 128–37.

Chen, Z. Y. *et al.* (2004) 'Silencing of episomal transgene expression by plasmid bacterial DNA elements in vivo.', *Gene therapy*, 11(10), pp. 856–864.

Conese, M., Auriche, C. and Ascenzioni, F. (2004) 'Gene therapy progress and prospects: episomally maintained self-replicating systems.', *Gene therapy*. Nature Publishing Group, 11(24), pp. 1735–41.

Costantini, M. (2006) ‘An isochore map of human chromosomes’, *Genome Research*, 16(4), pp. 536–541.

Van Craenenbroeck, K., Vanhoenacker, P. and Haegeman, G. (2000) ‘Episomal vectors for gene expression in mammalian cells.’, *European journal of biochemistry*, 267(18), pp. 5665–78.

Cremer, T. and Cremer, C. (2001) ‘Chromosome territories, nuclear architecture and gene regulation in mammalian cells.’, *Nature reviews. Genetics*, 2(4), pp. 292–301.

Dixon, J. R. *et al.* (2012) ‘Topological domains in mammalian genomes identified by analysis of chromatin interactions.’, *Nature*. Nature Publishing Group, 485(7398), pp. 376–80.

Dixon, J. R., Gorkin, D. U. and Ren, B. (2016) ‘Chromatin Domains: The Unit of Chromosome Organization’, *Molecular Cell*. Elsevier Inc., 62(5), pp. 668–680.

Dorer, D. R. and Henikoff, S. (1997) ‘Transgene repeat arrays interact with distant heterochromatin and cause silencing in cis and trans.’, *Genetics*, 147(3), pp. 1181–90.

Ebersole, T. A. (2002) ‘Mammalian artificial chromosome formation from circular alphoid input DNA does not require telomere repeats’, *Human Molecular Genetics*, 9(11), pp. 1623–1631.

Ehrhardt, A. *et al.* (2008) ‘Episomal vectors for gene therapy.’, *Current gene therapy*, 8(3), pp. 147–61.

Emery, D. W. *et al.* (2000) ‘A chromatin insulator protects retrovirus vectors from chromosomal position effects’, *Proceedings of the National Academy of Sciences*, 97(16), pp. 9150–9155.

Fournier, R. E. and Ruddle, F. H. (1977) ‘Microcell-mediated transfer of murine chromosomes into mouse, Chinese hamster, and human somatic cells.’, *Proceedings of the National Academy of Sciences of the United States of America*, 74(1), pp. 319–23.

Frappier, L. (2012) ‘Contributions of Epstein-Barr nuclear antigen 1 (EBNA1) to cell immortalization and survival.’, *Viruses*, 4(9), pp. 1537–47.

Gierman, H. J. *et al.* (2007) ‘Domain-wide regulation of gene expression in the human genome’, *Genome Research*, 17(9), pp. 1286–1295.

Girton, J. R. and Johansen, K. M. (2008) ‘Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila.’, *Advances in genetics*, 61(07), pp. 1–43.

Grosveld, F. *et al.* (1987) ‘Position-independent, high-level expression of the human beta-globin gene in transgenic mice.’, *Cell*, 51(6), pp. 975–85.

Guelen, L. *et al.* (2008) 'Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.', *Nature*, 453(7197), pp. 948–51.

Guy, L. G. *et al.* (1996) 'The beta-globin locus control region enhances transcription of but does not confer position-independent expression onto the lacZ gene in transgenic mice.', *The EMBO journal*, 15(14), pp. 3713–21.

Harrington, J. J. *et al.* (1997) 'Formation of de novo centromeres and construction of first-generation human artificial microchromosomes.', *Nature genetics*, 15(4), pp. 345–55.

Hiratsuka, M. *et al.* (2011) 'Integration-free iPS cells engineered using human artificial chromosome vectors.', *PloS one*, 6(10), p. e25961.

Hong, K., Sherley, J. and Lauffenburger, D. A. (2001) 'Methylation of episomal plasmids as a barrier to transient gene expression via a synthetic delivery vector', *Biomolecular Engineering*, 18(4), pp. 185–192.

Jenke, A. C. W. *et al.* (2004) 'Nuclear scaffold/matrix attached region modules linked to a transcription unit are sufficient for replication and maintenance of a mammalian episome', *Proceedings of the National Academy of Sciences*, 101(31), pp. 11322–11327.

Kazuki, Y. *et al.* (2010) 'Complete genetic correction of ips cells from Duchenne muscular dystrophy.', *Molecular therapy : the journal of the American Society of Gene Therapy*, 18(2), pp. 386–93.

Kazuki, Y. and Oshimura, M. (2011) 'Human artificial chromosomes for gene delivery and the development of animal models.', *Molecular therapy : the journal of the American Society of Gene Therapy*, 19(9), pp. 1591–601.

Kim, J.-H. *et al.* (2011) 'Human artificial chromosome (HAC) vector with a conditional centromere for correction of genetic deficiencies in human cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp. 20048–53.

Kim, J. M. *et al.* (2004) 'Improved recombinant gene expression in CHO cells using matrix attachment regions', *Journal of biotechnology*, 107(2), pp. 95–105.

Kouprina, N. *et al.* (2013) 'A new generation of human artificial chromosomes for functional genomics and gene therapy.', *Cellular and molecular life sciences : CMLS*, 70(7), pp. 1135–48.

Kwaks, T. H. J. *et al.* (2003) 'Identification of anti-repressor elements that confer high and stable protein production in mammalian cells.', *Nature biotechnology*, 21(5), pp. 553–8.

Liskovykh, M. *et al.* (2016) 'Moving toward a higher efficiency of microcell-mediated chromosome transfer.', *Molecular therapy. Methods & clinical development*, 3, p. 16043.

Lufino, M. M. P., Edser, P. A. H. and Wade-Martins, R. (2008) 'Advances in high-capacity extrachromosomal vector technology: Episomal maintenance, vector delivery, and transgene expression', *Molecular Therapy*. The American Society of Gene Therapy, 16(9), pp. 1525–1538.

Mejía, J. E. *et al.* (2001) 'Functional complementation of a genetic deficiency with human artificial chromosomes.', *American journal of human genetics*, 69(2), pp. 315–26.

Milanesi, G. *et al.* (1984) 'BK virus-plasmid expression vector that persists episomally in human cells and shuttles into Escherichia coli.', *Molecular and cellular biology*, 4(8), pp. 1551–60.

Mills, W. *et al.* (1999) 'Generation of an approximately 2.4 Mb human X centromere-based minichromosome by targeted telomere-associated chromosome fragmentation in DT40.', *Human molecular genetics*, 8(5), pp. 751–61.

Müller-Kuller, U. *et al.* (2015) 'A minimal ubiquitous chromatin opening element (UCOE) effectively prevents silencing of juxtaposed heterologous promoters by epigenetic remodeling in multipotent and pluripotent stem cells.', *Nucleic acids research*, 43(3), pp. 1577–92.

Nanbo, A., Sugden, A. and Sugden, B. (2007) 'The coupling of synthesis and partitioning of EBV's plasmid replicon is revealed in live cells.', *The EMBO journal*, 26(19), pp. 4252–62.

Ou, H. D. *et al.* (2017) 'ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells', *Science*, 357(6349).

Parada, L. A. *et al.* (2002) 'Conservation of relative chromosome positioning in normal and cancer cells.', *Current biology : CB*, 12(19), pp. 1692–7.

Pauler, F. M. *et al.* (2009) 'H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome.', *Genome research*, 19(2), pp. 221–33.

Peric-Hupkes, D. *et al.* (2010) 'Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.', *Molecular cell*, 38(4), pp. 603–13.

Phi-Van, L. *et al.* (1990) 'The chicken lysozyme 5' matrix attachment region increases transcription from a heterologous promoter in heterologous cells and dampens position effects on the expression of transfected genes.', *Molecular and cellular biology*, 10(5), pp. 2302–7.

Piechaczek, C. *et al.* (1999) 'A vector based on the SV40 origin of replication and chromosomal S/MARs replicates episomally in CHO cells', *Nucleic Acids Research*, 27(2), pp. 426–428.

Piirsoo, M. *et al.* (1996) 'Cis and trans requirements for stable episomal maintenance of the BPV-1 replicator.', *The EMBO journal*, 15(1), pp. 1–11.

Pikaart, M. J., Recillas-Targa, F. and Felsenfeld, G. (1998) 'Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators', *Genes & Development*, 12(18), pp. 2852–2862.

Ramunas, J. *et al.* (2007) 'Real-time fluorescence tracking of dynamic transgene variegation in stem cells.', *Molecular therapy : the journal of the American Society of Gene Therapy*, 15(4), pp. 810–7.

Rao, S. S. P. *et al.* (2014) 'A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping', *Cell*. Elsevier Inc., 159(7), pp. 1665–1680.

Rawlins, D. R. *et al.* (1985) 'Sequence-specific DNA binding of the Epstein-Barr virus nuclear antigen (EBNA-1) to clustered sites in the plasmid maintenance region.', *Cell*, 42(3), pp. 859–68.

Robertson, G. *et al.* (1995) 'Position-dependent variegation of globin transgene expression in mice.', *Proceedings of the National Academy of Sciences of the United States of America*, 92(12), pp. 5371–5.

Shopland, L. S. *et al.* (2006) 'Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence.', *The Journal of cell biology*, 174(1), pp. 27–38.

Stehle, I. M. *et al.* (2007) 'Establishment and mitotic stability of an extra-chromosomal mammalian replicon', *BMC Cell Biology*, 8, pp. 1–12.

Takahashi, Y. *et al.* (2010) 'Development of evaluation system for bioactive substances using human artificial chromosome-mediated osteocalcin gene expression.', *Journal of biochemistry*, 148(1), pp. 29–34.

Wang, S. *et al.* (2016) 'Spatial organization of chromatin domains and compartments in single chromosomes.', *Science (New York, N.Y.)*, 353(6299), pp. 598–602.

Wen, B. *et al.* (2009) 'Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells.', *Nature genetics*, 41(2), pp. 246–50.

Williams, S. *et al.* (2005) 'CpG-island fragments from the HNRPA2B1/CBX3 genomic locus reduce silencing and enhance transgene expression from the hCMV promoter/enhancer in mammalian cells.', *BMC biotechnology*, 5, p. 17.

Yates, J. L., Warren, N. and Sugden, B. (1985) 'Stable replication of plasmids derived from Epstein-Barr virus in various mammalian cells.', *Nature*, 313(6005), pp. 812–5.

Yu, J. *et al.* (2009) ‘Human induced pluripotent stem cells free of vector and transgene sequences.’, *Science (New York, N.Y.)*, 324(5928), pp. 797–801.

Zboray, K. *et al.* (2015) ‘Heterologous protein production using euchromatin-containing expression vectors in mammalian cells.’, *Nucleic acids research*, 43(16), p. e102.

Zhu, J. *et al.* (2013) ‘Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues.’, *Cell*. Elsevier Inc.

CHAPTER 2: VERSATILE MULTI-TRANSGENE EXPRESSION USING IMPROVED BAC TG-EMBED TOOLKIT, NOVEL BAC EPISOMES, AND BAC-MAGIC

ABSTRACT

Achieving reproducible, stable, and high-level transgene expression in mammalian cells remains problematic. Previously, we attained copy-number-dependent, chromosome-position-independent expression of reporter minigenes by embedding them within a BAC containing the mouse *Msh3-Dhfr* locus (DHFR BAC). Here we extend this "BAC TG-EMBED" approach. First, we report a toolkit of endogenous promoters capable of driving transgene expression over a 0.01-5 fold expression range relative to the CMV promoter, allowing fine-tuning of relative expression levels of multiple reporter genes expressed on a single BAC. Second, we show small variability in both the expression level and long-term expression stability of a reporter gene embedded in BACs containing either transcriptionally active or inactive genomic regions, making choice of BACs more flexible. Third, we describe an intriguing phenomenon in which BAC transgenes are maintained as episomes in a large fraction of stably selected clones. Finally, we demonstrate the utility of BAC TG-EMBED by simultaneously labeling three nuclear compartments in 94% of stable clones using a multi-reporter DHFR BAC, constructed with a combination of synthetic biology and BAC recombineering tools. Our extended BAC TG-EMBED method provides a versatile platform for achieving reproducible, stable simultaneous expression of multiple transgenes maintained either as episomes or stably integrated copies.

INTRODUCTION

Transgene expression has been widely used in both basic research and biotechnology. Applications of transgene expression range from the elucidation of gene function by ectopic expression of selected transgenes, to the expression of transgenes for gene therapy, and to the overexpression of genes for production of biopharmaceuticals (Wurm, 2004; Glover, Lipps and Jans, 2005; Prelich, 2012; Zhu, 2012; Walsh, 2018). Examples of such applications include the expression of multiple fluorescent proteins for live-cell imaging (Rizzo, Davidson and Piston, 2009), the expression of the four or more Yamanaka transcription factors for efficient generation of induced pluripotent stem (iPS) cells (Takahashi and Yamanaka, 2006), and the expression of multiple proteins for reconstitution of protein complexes (Machida *et al.*, 2012).

Despite the currently widespread use of transgene expression, most transgene expression systems still suffer from serious experimental limitations. Plasmid-, lentivirus- and transposon- based systems, all still show varying degrees of chromosome position effects (Akhtar *et al.*, 2013; Chen *et al.*, 2013) and position effect variegation (PEV) (Karpen, 1994; Robertson *et al.*, 1995; Ramunas *et al.*, 2007; Girton and Johansen, 2008; Tchasovnikarova *et al.*, 2015). Moreover, foreign sequences by themselves are targets for epigenetic silencing (Scrable and Stambrook, 1999; He, Yang and Chang, 2005; Suzuki, Kasai and Saeki, 2006; Minoguchi and Iba, 2008), and transgene concatamers can induce the formation of heterochromatin (Dorer and Henikoff, 1997; Garrick *et al.*, 1998). Together these transgene silencing mechanisms result in unpredictable transgene expression levels that do not correlate with copy number and are

unstable with long-term culture or changes in the cell physiological or differentiated state (Laker *et al.*, 1998; Hotta and Ellis, 2008; Herbst *et al.*, 2012).

Such limitations are compounded when the simultaneous and reproducible expression of multiple transgenes is required. For example, a common application in the emerging field of synthetic biology is the design of novel gene circuits, involving the expression of multiple proteins, in many cases at precise relative levels (Brophy and Voigt, 2014). While this approach has worked well in prokaryotes and yeast, it has been difficult to implement in mammalian cells due to the lack of suitable multi-transgene expression methods which overcome chromosome position effects and allow expression of different transgenes at reproducible relative levels.

A commonly used approach to countering this transgene-silencing phenomenon has been through the inclusion of *cis*-elements. These include insulators (Pikaart, Recillas-Targa and Felsenfeld, 1998; Emery *et al.*, 2000), locus control regions (LCRs) (Grosveld *et al.*, 1987; Guy *et al.*, 1996), scaffold/matrix attachment regions (S/MARs) (Phi-Van *et al.*, 1990; Kim *et al.*, 2004), ubiquitous chromatin opening elements (UCOEs) (Williams *et al.*, 2005; Müller-Kuller *et al.*, 2015) and anti-repressors (Kwaks *et al.*, 2003); some of these regulatory elements have context-dependent and/or vector dependent activity. While these *cis*-elements improve transgene expression to varying degrees, they are insufficient for chromosome-position independent, copy-number-dependent transgene expression (Guy *et al.*, 1996; Bharadwaj *et al.*, 2003; Truffinet *et al.*, 2005; Grandchamp *et al.*, 2011).

Additionally, in some transgene expression applications the ability to avoid transgene chromosomal integration and eventually eliminate these transgenes from the

cells is highly desirable. Both viral-sequence based and non-viral, pEPI based episomal vectors have been developed (Van Craenenbroeck, Vanhoenacker and Haegeman, 2000; Conese, Auriche and Ascenzioni, 2004; Ehrhardt *et al.*, 2008; Lufino, Edser and Wade-Martins, 2008). Viral-based vectors have the potential of causing transformation of the transfected cells (Frappier, 2012), while pEPI-like vectors, containing a S/MAR sequence immediately downstream of an active transcription unit, are mitotically stable without selection (Piechaczek *et al.*, 1999; Baiker *et al.*, 2000; Jenke *et al.*, 2004; Stehle *et al.*, 2007; Argyros *et al.*, 2008), and thus cannot be removed from the cells. Moreover, transgenes on these episomal vectors are still subject to silencing (Tessadori *et al.*, 2010), possibly due to the prokaryotic or viral sequences on these vectors (Chen *et al.*, 2004; Riu *et al.*, 2007).

Bacterial artificial chromosomes (BACs) carrying ~100-200 kb mammalian genomic DNA inserts harbor most of the *cis*-regulatory sequences required for expression of the endogenous genes contained within these genomic inserts. Previously we demonstrated how embedding minigene constructs at different locations within the DHFR BAC provided reproducible expression of single or multiple reporter genes independent of the chromosome integration site (Bian and Belmont, 2010). Similar approaches were used by other labs for high-level recombinant protein production (Blaas *et al.*, 2009; Zboray *et al.*, 2015). Recently, our lab demonstrated stable transgene expression after cell-cycle arrest or after terminal cell differentiation, using the BAC-TG EMBED approach (Chaturvedi *et al.*, 2018). All of these studies tested only BACs containing actively transcribed regions, based on the hypothesis that the expression level of the transgenes inserted into the BACs was determined by the chromatin environments

reconstituted by the genomic inserts within the BACs. Indeed, because of this assumption, previous studies have specifically targeted the inserted transgenes to transcription units and even exons (Blaas *et al.*, 2009; Bian and Belmont, 2010; Zboray *et al.*, 2015).

However, this hypothesis has not been tested. Moreover, overexpression from the genes on the BAC genomic inserts might change the properties of the transfected cells, or interfere with other assays of a study. Thus, BACs with no transcription units would be more desirable. Another improvement over our previous BAC TG-EMBED system (Bian and Belmont, 2010; Chaturvedi *et al.*, 2018) would be a toolkit of endogenous promoters capable of driving transgene expression over a wide range of expression levels. Viral promoters, including the CMV promoter we used previously, are known to be prone to epigenetic silencing (Fitzsimons, Bland and During, 2002; Brooks *et al.*, 2004), while most previously used endogenous and synthetic promoters were selected for their strength (Hong *et al.*, 2007; Qin *et al.*, 2010; Chen *et al.*, 2011; Zboray *et al.*, 2015). While high-level transgene expression is preferable in applications calling for overexpression, a low or near-physiological expression is important for many other applications, including gene therapy. Additionally, multiple transgenes may need to be expressed simultaneously but at reproducible differential levels.

Here we describe further extensions to the BAC TG-EMBED method that together provide a more versatile BC TG-EMBED toolkit for a range of future potential applications. First, we describe a toolkit of endogenous promoters that can drive transgene expression at reproducible relative levels over a 500-fold range. Second, we show that multiple BAC scaffolds can be used to drive sustained high-level transgene

expression driven by the UBC promoter without selection for up to 12 weeks, including BAC scaffolds containing no active transcription units. Third, we describe an episomal version of BAC TG-EMBED, where BAC transgenes form circular, ~1 Mb episomes and can be eliminated from the cells by removing selection. Fourth, we developed a "BAC-MAGIC" (**BAC**-**M**odular **A**ssembly of **G**enomic loci **I**nterspersed **C**assettes) to more rapidly assemble BACs containing multiple transgene expression cassettes. Finally, as a proof-of-principle demonstration of our new, more versatile BAC TG-EMBED toolkit, we demonstrate simultaneous expression of fluorescently tagged proteins labeling three different nuclear compartments, achieving >90% optimally labeled cell clones after a single, stable transfection.

RESULTS

Overview of BAC TG-EMBED toolkit development

We previously demonstrated the feasibility of the BAC TG-EMBED approach using both the DHFR BAC (Bian and Belmont, 2010) and a BAC containing the human GAPDH gene locus (GAPDH BAC) (Chaturvedi *et al.*, 2018). We set out to extend this BAC TG-EMBED methodology in two new directions (Figure 2.1).

First, to better control transgene expression and to be able to express multiple transgenes at reproducible expression ratios, we explored a set of constitutive promoters with various strengths for transgene expression. Testing each promoter with each BAC scaffold would have generated too large a number of possible combinations. We therefore decided to test a number of different promoters with the original DHFR BAC.

Second, we used one specific reporter gene construct to survey the effect of different BAC scaffolds on reporter gene expression. Previous applications of BAC TG-EMBED used BAC scaffolds containing multiple endogenous genes which would also be expressed in addition to added transgenes. Here we compared BAC scaffolds containing expressed genes with BAC scaffolds from gene deserts or regions containing silenced genes. We assayed the level, stability, and reproducibility of the embedded reporter gene expression when inserted into different BAC scaffolds to identify optimal BAC scaffolds for the BAC TG-EMBED system.

A toolset of 7 endogenous promoters for tuning relative transgene expression levels

We selected 7 endogenous promoters to test, either because of their known ability and use to drive transgene expression in a range of cell types (EEF1α, UBC) (Mizushima and Nagata, 1990; Schorpp *et al.*, 1996; Lois *et al.*, 2002; Hong *et al.*, 2007), or because these promoters were from housekeeping genes (RPL32, PPIA, B2M, RPS3A, GUSB) known to be expressed uniformly across a wide range of tissue types (de Jonge *et al.*, 2007; Hong Cai *et al.*, 2007; Zhu *et al.*, 2008; She *et al.*, 2009). We amplified 1-3 kb of regulatory regions upstream of the transcription start sites of these genes using either human genomic DNA as a template or, for the UBC promoter, using the pUGG plasmid (Chaturvedi *et al.*, 2018).

To assay relative promoter strength, we used the two-minigene reporter system developed in our previous study in which we compared expression of CMV-driven EGFP and mRFP minigenes inserted in the same mouse DHFR BAC scaffold (Bian and Belmont, 2010). We previously showed that the mRFP minigene reporter expression

varied less than or equal to 2.4-fold when the mRFP reporter was inserted at 6 different positions ranging 3-80 kb away from the EGFP reporter gene location on the same BAC (Bian and Belmont, 2010). To compare relative promoter strengths, we fixed the insertion positions of mRFP and EGFP, and measured the relative fluorescence levels of mRFP and EGFP when they were both driven by the CMV promoter versus when the mRFP reporter was driven by an endogenous promoter (Figure 2.1). Thus our assay measured the strength of different endogenous promoters relative to the viral CMV promoter.

For this assay, the EGFP reporter minigene was inserted 26kb downstream of the Msh3 transcription start site (Bian and Belmont, 2010) (Figure 2.2a). PCR-amplified promoters from 7 different housekeeping genes were cloned upstream of the mRFP expression cassette (Figure 2.2b), and then this mRFP expression cassette was introduced 121 kb downstream of the Msh3 transcription start site by BAC recombineering (Figure 2.2a), generating the dual reporter DHFR BAC. As a control, we used the dual reporter BAC previously constructed (Bian and Belmont, 2010) in which the same mRFP cassette driven by the CMV promoter was inserted at this same location 121 kb downstream of the Msh3 start site.

Mouse NIH 3T3 fibroblasts were then stably transfected with these modified BAC constructs. After dual selection with G418 and Zeocin for two weeks, mixed populations of stable clones carrying the BAC transgenes were analyzed by flow cytometry to measure the relative expression ratio of mRFP and EGFP (Figure 2.2c). Fluorescent beads were used as an invariant fluorescence standard to calibrate the flow cytometer intensity outputs. The ratio of mRFP to EGFP expression was then normalized by the ratio observed with the original dual-reporter BAC construct in which both

reporters were driven by CMV promoter, providing the endogenous promoter strength relative to the CMV promoter.

We observed an overall variation in promoter strength of over 500-fold, ranging from the 4-5 fold relative promoter strength of the RPL32 and EEF1α promoters to the 0.01-fold relative promoter strength for the GUSB promoter as compared to the CMV promoter (Figure 2.2d).

Reporter gene expression as a function of transcriptionally active and inactive BAC scaffolds

To find the best BAC scaffold for the BAC TG-EMBED system, we tested BAC scaffolds from both actively transcribed regions and regions containing silenced genes or no genes. Specifically, we measured the expression as a function of copy number of one specific reporter gene construct inserted into these BAC scaffolds. Previous applications of BAC TG-EMBED showed a linear relationship between copy number and expression level, largely independent of the chromosome integration site, demonstrating copy-number dependent, position independent transgene expression (Bian and Belmont, 2010; Chaturvedi *et al.*, 2018). For active chromosomal regions, we chose the RP11-138I1 BAC containing the human ubiquitin B gene locus (UBB BAC), the RP23-401D9 BAC containing the "safe-haven" mouse *Rosa26* genetrap locus (ROSA BAC) (Zambrowicz *et al.*, 1997), and the CITB-057L22 BAC carrying the mouse Dhfr gene locus (DHFR BAC). For inactive chromosomal regions, we chose the CTD-2207K13 BAC (2207K13 BAC) that contains no known gene or regulatory element from a gene-desert region from the human genome, and the CTD-2643I7 (HBB BAC) containing the human HBB gene

locus and multiple olfactory genes, all of which are transcriptionally silenced in fibroblasts (Bertulat *et al.*, 2012).

We selected the UBC promoter for this reporter gene cassette as this promoter had previously been shown to drive high expression across multiple cell types (Lois *et al.*, 2002); in our dual reporter system the UBC promoter was 2.6-fold stronger than the CMV promoter (Figure 2.2d). Moreover, to eliminate any possible transcriptional interference from closely spaced reporter and selectable marker minigenes and to minimize any epigenetic silencing arising from DNA methylation of this reporter gene-selectable marker construct, we used a commercially available GFP-ZeoR fusion protein gene construct in which all CpG dinucleotides had been removed and replaced by synonymous codons (Figure 2.3a).

We inserted this UBC-GFP-ZeoR reporter gene construct into different BAC scaffolds by BAC recombineering, using *galK* for positive/negative selection (Warming *et al.*, 2005; Khanna *et al.*, 2013). To eliminate potential artifacts caused by proximity to active promoters, transcriptional start sites (TSS), or miRNA sequences, we chose insertion sites flanked on both sides by at least 5 kb free of such sequence elements (Figure 2.3b). The UBB, HBB, 2207K13, ROSA, DHFR BACs with the UBC-GFP-ZeoR reporter gene insertion were named as UBB-UG, HBB-UG, 2207K13-UG, ROSA-UG and DHFR-UG.

After transfection, multiple cell clones (n=20-40) carrying stably integrated BAC arrays were selected for Zeocin resistance and analyzed for reporter gene expression by flow cytometry, using untransfected NIH 3T3 cells to determine background, autofluorescence levels. For each cell clone, we used flow cytometry to measure the

mean GFP reporter expression and qPCR to measure reporter gene copy number. These cell clones showed GFP fluorescence mean levels ranging from 10-1000 fold higher than the background autofluorescence.

Our original working hypothesis predicted that the BAC TG-EMBED reporter expression should be uniform in all cells of the same clone. Also, we expected to see a linear relationship between mean reporter gene fluorescence and number of BAC copies, signifying a copy-number-dependent, position independent expression. Furthermore, we expected that the slope of this linear relationship would be higher for BAC scaffolds expected to reconstitute an active chromatin environment permissive for transgene expression as compared to BAC scaffolds expected to reconstitute a more condensed, inactive chromatin environment (Figure 2.1). In contrast, we expected that the reporter gene cassette transfected without any BAC scaffold would show clonal expression levels that poorly correlated with reporter gene copy number (copy-number-independent expression).

Unexpectedly, the stable cell clones we isolated showed two distinct types of population expression profiles- uniform versus heterogeneous. Uniform clones showed single, relatively narrow expression peaks in the flow cytometry histograms, with more than 90% of the cells showing GFP fluorescence varying only over a 10-fold intensity range (Figure 2.3c, left). Heterogeneous clones instead showed two peaks with a range of GFP expression varying ~1000-fold, with the lower GFP intensity peak overlapping with the autofluorescence distribution of control cells (Figure 2.3c, right). We had not previously observed such heterogeneous expression profile using our original DHFR BAC containing the CMV-driven mRFP alone or both the CMV-driven EGFP and CMV-

driven mRFP reporter genes (Bian and Belmont, 2010). However, we had observed ~80% uniform clones for a GAPDH BAC scaffold with the UBC-GFP-ZeoR reporter gene inserted (Chaturvedi *et al.*, 2018). The percentage of clones showing such heterogeneous expression varied from 58% to 83% for the 5 BAC scaffolds surveyed here (Table 2.1). No similar heterogeneous expression profile was observed when the reporter gene construct was transfected by itself (Table 2.1).

As expected, the control transfection of the reporter gene cassette by itself resulted in copy-number-independent expression of the reporter gene (Figure 2.3d, R^2=0.09), while the reporter gene embedded within the BACs yielded a linear relationship between reporter gene fluorescence for both uniform (black) and heterogeneous (red) BAC transgene clones (Figure 2.3d, R^2=0.561 to 0.914).

Surprisingly, we observed no more than a 4-fold variation in expression per copy number among the 5 different BAC scaffolds tested, with no obvious relationship between the observed slope and the type of BAC scaffold (Figure 2.3d). Although the transcriptionally active DHFR BAC produced the highest slope, the transcriptionally inactive HBB BAC and the 2207K13 BAC containing DNA from a gene desert produced the second and third highest slopes, while the BAC containing DNA from the "safe haven" mouse *Rosa26* locus produced the lowest slope.

Overall, these results show that for this UBC-GFP-ZeoR reporter gene, high-level, copy-number-dependent transgene expression using the BAC TG-EMBED method does not require BACs containing active, housekeeping genomic regions, but can also be obtained from a wide range of BAC genomic DNA inserts, including gene-desert regions. This means BAC TG-EMBED can be used to drive expression of only the transgenes

added to the BAC scaffold, without overexpression of the genes contained within the BAC scaffold.

Temporal stability of BAC-embedded reporter gene expression in uniform cell clones

We previously showed that the BAC TG-EMBED method provided long-term stability of transgene expression in the presence of continued drug selection (Bian and Belmont, 2010). However, in the absence of drug selection we observed a 30-80% drop in expression over several months of cell passaging without any apparent drop in the integrated BAC copy number (Bian and Belmont, 2010).

Here we determined the long-term stability of the UBC-GFP-ZeoR reporter gene expression for both uniform and heterogeneous clones for four different BAC scaffolds. Individual clones for each BAC scaffold (3 uniform and 2 heterogeneous for ROSA-UG BAC, 7 uniform and 4 heterogeneous for 2207K13-UG BAC, 8 uniform and 3 heterogeneous for UBB-UG BAC, and 3 uniform and 3 heterogeneous for DHFR-UG BAC) were passaged up to three months in the absence or presence of drug selection and analyzed for reporter gene fluorescence at regular intervals after removal of drug selection.

With the exception of a small number of apparent fluctuations possibly related to transient changes in culture conditions, clones with uniform reporter gene expression showed no significant change either in the mean fluorescence values (Figure 2.4a) or in the distribution of fluorescence among the same clones (Figure 2.4b and Figure A.1) over time in the absence of selection for all four BAC scaffolds tested. In the presence of continued selection, uniform clones containing DHFR-UG or ROSA-UG BACs showed

no significant reporter gene expression change, while an ~50% or 100% increase was observed for the UBB-UG or 2207K13-UG BAC clones, respectively (Figure 2.4a). No changes in estimated BAC copy number based on qPCR measurement were observed for any of these clones during this time series. This suggests that epigenetic changes driven by selection pressure may be responsible for these small increases in reporter gene expression.

Notably, in the absence of selection, heterogeneous clones for all tested BAC scaffolds showed a significant and progressive loss of reporter gene expression over time. This led to a significant fraction of cells showing autofluorescence levels of fluorescence by the end of the experiment (Figure 2.4a). Reporter gene expression-level became progressively more homogenous, but at lower fluorescence levels (Figure 2.4b and Figure A.1). With selection, UBB-UG and DHFR-UG BAC heterogeneous clones showed a 1.6 to 3-fold increase in reporter gene expression, respectively, while the other BAC scaffold heterogeneous clones showed no significant changes (Figure 2.4a).

BAC transgenes are maintained as episomes in heterogeneous clones

In our previous work, all stable cell clones obtained after BAC transfection and drug selection contained single BAC copies or multi-copy BAC arrays that had integrated into endogenous chromosomes (Hu *et al.*, 2009; Bian and Belmont, 2010; Hu, Plutz and Belmont, 2010; Sinclair *et al.*, 2010; Bian *et al.*, 2013) consistent with similar results from numerous laboratories. Thus, we initially assumed that the broad distribution of reporter gene fluorescence observed in heterogeneous cell clones was due to position effect variegation (PEV) of the BAC TG-EMBED reporter genes. We hypothesized that

integrations into some chromosome integration sites led to uniformly-expressing clones, while integration into other chromosome sites prone to PEV led to heterogeneous clones with variegated transgene expression.

However, the observation of a progressive loss over time of reporter gene expression for all heterogeneous clones led us to question the genome stability of the BAC transgenes in these clones. To test the relationship between changes in reporter gene expression and BAC copy number, we first sorted cells from the heterogeneous DHFR-UG-s3 cell clone by fluorescence-activated cell sorting (FACS), using a narrow sorting-window centered around the GFP peak fluorescence level (Figure 2.5a). After cell-sorting, with drug selection the original heterogeneous reporter gene expression distribution reestablished itself within one week of culture (Figure 2.5b). We then resorted cells showing different levels of GFP fluorescence using four narrow fluorescence windows P1, P2, P3, and P4 (Figure 2.5b), and then used qPCR to measure the BAC copy number in cells from each of these sorting windows. Plotting mean cell fluorescence intensity levels versus copy number for these clonal subpopulations yielded a strikingly linear relationship (R^2=0.99) (Figure 2.5c). Thus, the variable reporter gene expression level in this heterogeneous cell clone is the result of loss of BAC transgenes rather than chromosome PEV.

To identify the source of this BAC copy-number instability, we next used DNA FISH to visualize BAC transgenes within interphase nuclei and mitotic chromosome spreads. We compared the distribution of BAC transgenes within the heterogeneous clone, DHFR-UG-s3, versus a uniform clone, DHFR-UG-f3-15.

DNA FISH suggested that whereas the uniform clone contained cells with an integrated DHFR BAC array, the heterogeneous clone contained cells in which the DHFR BAC was present as episomes. Specifically, interphase FISH against the DHFR BAC in the heterogeneous clone revealed multiple, noncontiguous, small spots distributed randomly throughout the nuclei (Figure 2.5d). The number of these spots was highly variable in different cell nuclei, suggesting unequal segregation of BAC transgenes. In contrast, most cells from the uniform clone showed just one large, fiber-like FISH spot per nucleus (Figure 2.5e). Moreover, FISH spots in mitotic spreads from the heterogeneous clone were either touching or spatially separated from the chromosomes (Figure 2.5f), whereas FISH spots in mitotic spreads from the uniform clone were always located within the chromosome (Figure 2.5g). The number of FISH spots per mitotic spread was highly variable in the heterogeneous clone, with each spot much smaller than the single FISH spot visualized within the mitotic chromosome from the uniform clone. Interestingly, the FISH spots in the heterogeneous clone mitotic spreads had weak DAPI staining, varying from slightly elevated over background to no difference from background (Figure 2.5h), suggesting these structures are much smaller than previously described double minute chromosomes (DMs) generated by gene amplification (Carroll *et al.*, 1988; Schwab and Amler, 1990; L'Abbate *et al.*, 2014).

Using DNA FISH of both interphase nuclei and mitotic spreads, we confirmed this finding of integrated BACs in uniformly expressing clones versus episomal BACs in heterogeneously expressing clones in additional cell clones carrying BAC transgenes based on three different BAC scaffolds (FigureA.2 and data not shown- 3 heterogeneous and >3 uniform clones for both DHFR and HBB BAC scaffolds).

Unequal segregation of these BAC episomes during cell division would explain the heterogeneity of BAC transgene copy number in the cell population of heterogeneous clones, leading to variability of reporter gene expression. Indeed, telophase cells from heterogeneous clones showed unequal numbers of FISH spots in the two daughter nuclei (Figure 2.5i). In the absence of continued drug selection, we would expect cells that have lost BAC transgenes will accumulate if there is any selective growth advantage for cells with fewer BAC copies.

BAC episomes are circular and ~1 Mb in size

We analyzed the average amount of DNA per BAC episome, using two independent methods- light microscopy and pulsed-field gel electrophoresis (PFGE). Both methods produced a similar estimate of ~800-1000 kb per BAC episome.

Using light microscopy, we measured the average DHFR BAC episome DAPI integrated staining intensity in mitotic spreads from cell clone DHFR-UG-s3 relative to the smallest mouse chromosome (chr19) with known DNA content of 61.4 Mbp (Figure 2.6a). This comparison produced an estimated mean episome size of 770 kb in this DHFR-UG-s3 clone.

Using PFGE, we observed that the BAC episomes were circular and estimated the modal BAC episome size to be ~900 kb and 1 Mbp for DHFR-UG and HBB-UG BAC episomes in cell clones DHFR-UG-s3 and HBB-UG-100d3, respectively. Two different cell clones, DHFR-UG-f3-1 and HBB-UG-fD2, were used as negative controls as they contained the same DHFR-UG or HBB-UG BAC DNA as the cell clones with episomes but the BAC DNA was integrated within endogenous mouse chromosomes. *E. coli*

strains containing the DHFR or HBB BACs were used as positive controls for detection of circular episomes.

Pulsed-field gels were analyzed by Southern blotting using pooled BAC DNA PCR products as the hybridization probes. Linear but not circular DNAs migrate in pulsed-field gels. Similar to the *E. coli* controls containing circular BACs, the Southern blot signals for the BAC DNA from the two clones containing episomes did not migrate out of the wells (Figure 2.6b and FigureA.3a-b), consistent with circular rather than linear BAC episomes.

To validate that the BAC episomes are really circular, and to estimate their size, the agarose-embedded DNA was digested using the ssDNA specific Nuclease S1 prior to PFGE and Southern blot hybridization. After removal of proteins, circular DNA episomes in both bacteria and mammalian cells are typically negatively supercoiled. This supercoiling generates torsional stress which is relieved by local formation of single-stranded regions. Thus, S1 nuclease has been used to cut these single-stranded regions and linearize circular DNA episomes (Barton, Harding and Zuccarelli, 1995; Walker, LeBlanc and Sikorska, 1997; Marasini and Fakhr, 2014). After S1 digestion, DNA from the cell clones carrying BAC episomes now showed DNA smears with peak intensities of ~900 kb and 1 Mb for the DHFR-UG-s3 and HBB-UG-100d3 cell lines, respectively (Figure 2.6b and Figure A.3a-b). In contrast, after S1 nuclease digestion, DNA from the integrated BAC clones showed signals within the wells and above 2 Mb, overlapping the smears of fragmented genomic DNA (Figure 2.6b and Figure A.3a-b). DNA of *E. coli* containing DHFR BAC and HBB BAC episomes produced bands at ~200-300 kb, in addition to signals in the wells (Figure 2.6b and Figure A.3a-b) after S1 nuclease

digestion. These estimated BAC sizes measured slightly larger than the actual BAC sizes (~200 kb), indicating there might be a slight overestimation of episome sizes using our PFGE running conditions.

BAC episomes contain no detectable host DNA as revealed by CNV analysis

The propagation of BAC transgenes as episomes was unexpected. A major question is whether these episomes consist solely of BAC DNA, or whether host DNA is also included and possibly required for episome propagation.

We first compared the estimated episome DNA content size with estimates of BAC copies per episome. BAC episome sizes estimated by either light microscopy or PFGE were approximately twice as large as predicted from qPCR BAC copy number estimates (Table 2). The estimated average BAC content per episome was 445 kb in DHFR-UG-s3 and 716 kb in HBB-UG-100d3.

The difference in the estimated BAC DNA content per episome and the average episome size is at most a few hundred kb, and may be accounted for by inaccuracies of the qPCR copy number, PFGE size estimation, and possible variation in sequence representation within the BAC episomes due to shearing of DNA and/or recombination during the transfection and creation of the BAC episomes.

Alternatively, this difference in episome size versus qPCR estimation of BAC copies per episome could also be caused by presence of host cell genomic DNA on the episomes. To search with higher sensitivity for the possible presence of host DNA within the episome, we performed Whole Genome Sequencing (WGS) based copy number variation (CNV) analysis of the two clones, DHFR-UG-s3 and HBB-UG-100d3.

Genomic regions present on the episomes would appear amplified in cells containing episomes (test sample), comparing to cells with no episomes (reference sample). Thus the ratio in the number of reads for a given bin between the test sample and the reference sample was calculated. To reduce noise, bins were merged into segments based on the $\log_2$(ratio), using a circular binary segmentation (CBS) algorithm (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007). The mean $\log_2$(ratio) of each segment was used to estimate the CNV of this segment in the test sample relative to the reference sample.

Mouse 3T3 cells show genomic instability; therefore we anticipated CNV between the parental cell line and individual clones. To reduce false-positives derived from CNV between different 3T3 clones, independent of episome content, we used cells with low reporter gene fluorescence sorted from the cell clone containing the episomal BAC transgenes (region L, Figure 2.6c-d, and Figure A.4a) as the reference sample. To further reduce false positives, we also imposed constraints that copy number increase for true positive regions should be reproducible between experimental replicates and correlate with episomal copy number. We calculated the estimated CNV in cells sorted with high (H2) reporter gene expression, using sorted cells with low (L) expression as the reference sample (H2, L, Figure 2.6c-d, and Figure A.4a). We also compared the estimated CNV in cells sorted with high reporter gene expression in an independent experiment (H1, Figure 2.6c-d, and Figure A.4a) with the estimated CNV from the first experiment. All samples were sequenced to ~2x coverage.

We used 3 and 30 kb bin sizes for analysis. To reduce noise, we excluded all bins in the test sample with zero reads (5.5% of total bins for 3 kb bin analysis and 3.5% for 30 kb bin analysis) plus extreme outlier bins, defined by the lower quantile minus 4 times

the interquantile distance, with unusually low read count (~0.5% of total bins for 3 kb bin analysis and 2.5% for 30 kb bin analysis) in the test sample before calculating ratios.

As a test of our analysis method, we compared the mean segment $\log_2$(ratio) of the BAC regions in H1 and H2, generated by the above analysis method, to the fold increase of BAC regions in H1 and H2 relative to L measured by qPCR. As expected, the results from the CNV and qPCR analysis were very similar (Figure A.5).

We were interested in asking whether a specific host DNA element was present on each episome copy present within a cell clone. We estimated that on average the sorted cells with high reporter gene expression had 15-20 episome copies per cell, depending on the cell clone, based on qPCR of BAC DNA sequences and the estimated number of BACs per episome (Table 2.3).

We estimated theoretical minimum copy number increase for episome-localizing host DNA (minimum increase) in the H1 and H2, based on BAC copy number measured by qPCR, and assuming the NIH 3T3 to be tetraploid and each episome to have the same host cell genomic DNA (Table 2.3). Segments with mean $\log_2$(ratio) equal to or greater than $\log_2$(minimum increase) in both H1 and H2 samples were selected as candidates for being on the episomes (Figure 2.6e).

Interestingly, all candidate segments identified belonged to the BAC regions, including the UBC-GFP-ZeoR and the BAC vector (Figure 2.6f-g, Figure A.4b-c and Figure A.6), and no other mouse genomic sequence satisfied all of the above conditions.

In conclusion, we could not detect host cell DNA reproducibly present on all episomal copies using bin sizes of either 3 or 30 kb. We therefore conclude BAC DNA itself is sufficient for the creation and propagation of these BAC episomes. We cannot

exclude the possibilities, however, that an unmappable, repetitive host DNA sequence is present on the episomes and confers their ability to propagate or that different host DNA sequences are present on each episome present within a single cell clone.

Multiple promoters added to BACs support formation of episomal BAC transgenes but only in certain cell lines

Because we did not observe episomal BAC transgenes in our original BAC-TG EMBED work using the CMV-mRFP-SV40-ZeoR reporter gene (Bian and Belmont, 2010), we hypothesized that addition of the UBC-GFP-ZeoR reporter gene might be responsible for BAC episome formation. Our dual-reporter assay showed that the UBC promoter was much stronger than the CMV promoter; therefore, we further hypothesized that promoter strength might correlate with the frequency of BAC episome formation.

To test this hypothesis, we isolated clones stably transfected with the dual-reporter DHFR BAC transgenes and examined reporter gene expression patterns in these clones by flow cytometry (Figure A.7a). As expected, no heterogeneously GFP/RFP expressing clones where observed when the mRFP reporter gene was driven by CMV promoter (n=13) or B2M promoter (n=6). In contrast, we observed ~70% or ~30% heterogeneously GFP/RFP expressing clones when the mRFP was driven by the EEF1a promoter (12/18) or the RPL32 promoter (10/29), respectively (Figure A.7a-b). We confirmed that BAC transgenes in these heterogeneously expressing clones were episomal using DNA FISH (Figure A.7c).

These results using human promoter sequences added to the BACs, did show a rough correlation of promoter strength with the frequency of clones containing episomes.

However, when we examined a series of DHFR BAC constructs, we instead observed clones with episomes using BAC transgenes containing the dual reporter, selectable marker CMV-mRFP-SV40-ZeoR reporter cassette. This included the identical DHFR BAC construct used in our previous BAC-TG EMBED work (Bian and Belmont, 2010), as well as various DHFR BAC deletions (Figure A.8a). All the DHFR BAC constructs contain LacO repeats, and a NIH 3T3 derived clone expressing EGFP-LacI was used for transfection, so that BAC transgenes could be observed directly in fixed cells. Although all clones showed a unimodal flow cytometry expression pattern, explaining why we did not observe this phenomenon previously, a large fraction (DHFR-c27: 1/2, DHFR-c27d2: 2/10, DHFR-c27d3crz: 2/16, DHFR-c27d4: 7/10) of clones showed episomal BAC transgenes (Figure A.8b-c).

Thus, the promoter used to drive reporter and/or selectable markers appears to determine not whether episomal BAC transgenes are established but rather whether a unimodal versus bimodal distribution is observed in cells containing these episomal BAC transgenes. The presence of strong promoters (UBC, EEF1a and RPL32) appears to allow the formation of bimodal distributions of reporter gene expression, possibly related to the balance between the degradation rate of the initially high levels of selectable marker versus the rate of loss of BAC transgene episomes during cell division.

Next, we tested whether BACs can form episomes in a different cell line other than mouse NIH 3T3 fibroblasts. Previously, we observed cell clones containing only integrated BAC transgenes in CHO (Hu *et al.*, 2009; Hu, Plutz and Belmont, 2010) and mouse ES cells (Sinclair *et al.*, 2010; Chaturvedi *et al.*, 2018). Reasoning that cancer cells with some level of genomic instability might be more prone to formation of BAC

episomes, we tested the human colorectal carcinoma epithelial cell line, HCT116, using the 2207K13-UG BAC which produced 79% episome clones in NIH 3T3 cells.

Four out of 32 stable clones showed a heterogeneous GFP distribution similar to that observed in NIH 3T3 episome clones, with a broad high fluorescent peak and a tail/secondary peak near the auto-fluorescence level (Figure A.9). However, none of these four clones showed episomal BAC transgenes by DNA FISH (Figure A.10a). Instead, most cells in each clone showed the same number (one or two) of spots, but these spots varied in size from cell to cell. Therefore, it appears that the broad GFP peaks in these four clones are due to some form of genomic instability leading to CNV of integrated transgene arrays. Interestingly, one clone, HCT116-k13_06, out of the 32, which had a single GFP peak, showed a small fraction of cells of with episomal BAC transgenes, in contrast to the vast majority of cells which contained integrated BAC transgenes (Figure A.10b). One out of 24 subclones of this HCT116-k13_06 clone, HCT116-k13_06-10, showed a similar mixed population with either integrated BACs or episomal BACs, similar to the parent clone HCT116-k13_06 (Figure A.10c). The low frequency of clones with episomal BACs, the variable size of the integrated BAC transgene arrays, and the co-existence of integrated and episomal BAC transgenes in the same cells and from the same clone suggests these episomes might arise from the well-known phenomenon of gene amplification (Carroll *et al.*, 1988; Schwab and Amler, 1990; L'Abbate *et al.*, 2014).

Similarly, a small percentage of clones carrying the GAPDH BAC in stable mouse ES cell colonies showed broad GFP expression peaks by flow cytometry, but FISH revealed this was due to variable size, integrated BAC transgene arrays (Appendix

B), due presumably to some type of CNV induced by genomic instability of these transgene arrays.

In conclusion, the high frequency establishment of BAC transgene episomes seen in mouse 3T3 cells does not appear to occur in either HCT116 or mouse ES cells, or at detectable frequency in CHO cells (Hu *et al.*, 2009; Hu, Plutz and Belmont, 2010; Sinclair *et al.*, 2010; Chaturvedi *et al.*, 2018).

Expression of multiple-reporters by BAC-MAGIC

As a proof-of-principle application of our improved toolkit for BAC TG-EMBED, we created a multi-transgene BAC to label simultaneously the nuclear lamina, nucleoli, and nuclear speckles with a single stable transfection. The original DHFR BAC was used for this multi-transgene expression. A SNAP-tagged Lamin B1 reporter mini-gene was used to label the nuclear lamina, a SNAP-tagged Fibrillarin the nucleoli, and an mCherry-Magoh the nuclear speckles. We used the RPL32 promoter to drive the expression of the SNAP-tagged Lamin B1, and a promoter of intermediate strength, PPIA, for the SNAP-tagged Fibrillarin and the mCherry-tagged Magoh, which are both abundant proteins.

Previously, we used random Tn5 transposition to introduce expression cassettes into BAC scaffolds (50), but this approach is limited in the number of serial insertions that can be made due to the remobilization of existing transposons, its requirement for multiple selectable markers, and the randomness of the insertion sites. Alternatively, BAC recombineering using antibiotic resistance genes as positive selectable markers have been used to insert expression cassettes into precise locations on the BACs. However, like transposition, this method relies on the availability of multiple selectable

markers and introduces unwanted selectable markers into the BACs. An alternative BAC recombineering scheme using cycles of *galK*-based positive selection to insert sequences followed by negative selection to remove *galK* have been used to make multiple BAC modifications without addition of unwanted selectable markers. However, the low efficiency of negative selection, due to a high background of competing, spontaneous deletions of mammalian DNA with its high repetitive DNA content, makes this approach quite time and labor intensive. Typically, one month is required for each cycle of insertion of DNA by positive selection, removal of the selectable marker by negative selection, and subsequent screening and testing of DNA from colonies that survive the negative selection to identify the small fraction of colonies containing the desired homology-driven, specific deletion of just the selectable marker.

To accelerate creation of BACs containing multiple transgene, we created a new BAC assembly approach, BAC MAGIC (**BAC**-**M**odular **A**ssembly of **G**enomic loci **I**nterspersed **C**assettes). BAC MAGIC combines the DNA assembler method in yeast (Shao, Zhao and Zhao, 2009; Shao and Zhao, 2012) and/or Gibson assembly (Gibson *et al.*, 2009) with traditional cloning methods to create a number of BAC recombination modules followed by sequential rounds of BAC recombineering in which one fragment is inserted using one selectable marker followed by addition of a new fragment overlapping the previous fragment using a second positive selectable marker which replaces the first (Richardson *et al.*, 2017). Each round of fragment insertion only requires ~ 1 week for transformation and screening of clones. In this way, 45 kb of the DHFR BAC was effectively reconstructed such that DHFR sequences remained but 3 fluorescent mini-gene expression cassettes were added, each spaced by ~10 kb of DHFR sequence (Figure

2.7a-b, Figure A.11). The large homologous sequences flanking each expression cassette reduces recombination between similar sequences in other expression cassettes already inserted into the BAC, increasing the efficiency of this overall approach.

We began the process using a DHFR BAC. After six rounds of BAC recombineering, we had created a BAC with four expression cassettes (Figure 2.7b): a SNAP-tagged Lamin B1 minigene, a SNAP-tagged Fibrillarin minigene, a mCherry-Magoh minigene, and a ZeoR selectable marker.

We tested simultaneous expression of the three reporters in 17 independent NIH 3T3 cell clones transfected with the multi-reporter BAC by examining fluorescence in fixed cells under a microscope (SNAP-tagged proteins were labeled with a Fluorescein conjugated SNAP tag substrate before fixation). We observed uniform expression of all the three reporters in 16/17 clones. The loss of SNAP-Lamin B1 expression in one of the clone (Cl#16) may be due to random breakage of the BAC during transfection, as PCR revealed the absence of this minigene from the cell clone. Similarly, 12/14 U2OS human osteosarcoma cell clones showed both SNAP-Lamin B1 and SNAP-Fibrillarin expression after transfection of a BAC containing only these two expression cassettes (data not shown).

Within individual cells, a linear correlation was observed between the integrated fluorescence intensity per cell of SNAP-tagged proteins Lamin B1 and Fibrillarin versus mCherry-Magoh in 4/4 representative NIH 3T3 clones (04, 08, 13 and 14, Figure 2.7c). Moreover, these fluorescently tagged proteins showed uniform rather than variegating expression in different cell nuclei of the same clone observed under the microscope (Figure 2.7d).

DISCUSSION

We previously demonstrated the utility of the BAC TG-EMBED method to achieve position-independent, copy-number-dependent, one-step transgene expression in mammalian cells (Bian and Belmont, 2010; Chaturvedi *et al.*, 2018). Here, we have extended the BAC TG-EMBED methodology through four new advances and provided a proof-of-principle demonstration of this new methodology by efficiently creating cell lines stably expressing uniform levels of three different fluorescently tagged proteins-Lamin B1, Fibrillarin, and Magoh in a single stable transfection.

First, we describe a toolkit of endogenous promoters providing an ~500-fold range in promoter strength varying from ~5 fold higher to ~100-fold weaker than the commonly used viral CMV promoter. As these promoters are from human genes shown to be expressed in a wide range of cell lines and tissues (Mizushima and Nagata, 1990; Schorpp *et al.*, 1996; Lois *et al.*, 2002; de Jonge *et al.*, 2007; Hong Cai *et al.*, 2007; Hong *et al.*, 2007; Zhu *et al.*, 2008; She *et al.*, 2009), we expect them to support transgene expression in most cell types and independent of cell proliferation or differentiation state. While most of the previous studies on transgene promoters focused on conventional, strong promoters (Hong *et al.*, 2007; Qin *et al.*, 2010; Chen *et al.*, 2011; Zboray *et al.*, 2015), we included weak promoters and promoters that have not been widely used in our survey. The weak promoters we identified, such as GUSB and RPS3A, could possibly replace the commonly used minimal promoters or prokaryotic inducible promoters where a sustained low-level of transgene expression is needed. Moreover, this wide range of promoter strengths allows reproducible expression of multiple transgenes over a wide range of relative expression levels from a single BAC scaffold.

Second, we show that with the UBC-GFP-ZeoR reporter gene, our BAC-TG EMBED system achieved stable reporter gene expression of integrated BAC transgenes for several months in the absence of drug selection. This is an improvement over the 30-80% drop in expression observed originally with the CMV-mRFP-SV40-ZeoR reporter gene (Bian and Belmont, 2010). Both the UBC promoter and the CpG free GFP-ZeoR gene body could be contributing to this improvement.

Third, we show that at least with the UBC-GFP-ZeoR expression cassette, our BAC TG-EMBED system is not dependent on BAC scaffolds containing active DNA genomic regions but also works with BAC scaffolds containing silenced DNA genomic regions as well as gene deserts. UBC may represent a member of a class of active, house-keeping gene promoters that is relatively insensitive to chromosome position effects. This allows choice of a BAC scaffold for the BAC TG-EMBED method that will not co-express any genes other than the introduced transgene cassettes. In contrast, both our previous BAC TG-EMBED studies (Bian and Belmont, 2010; Chaturvedi *et al.*, 2018) and similar work from other laboratories (Blaas *et al.*, 2009; Zboray *et al.*, 2015), used only BACs containing highly-transcribed house-keeping genes, due to the assumption that either an active chromatin region or active 5' *cis*-regulatory regions would be required for creating a transcriptionally permissive environment for transgene expression. Integration of the UBC-GFP-ZeoR reporter gene into the BAC was required for position-independent, copy-number dependent expression, as its expression was copy-number independent when the same UBC-GFP-ZeoR reporter gene was stably transfected by itself into cells. The expression levels of this UBC-GFP-ZeoR were similar, per copy

number, in cell clones with episomal BAC transgenes to levels in clones with integrated BACs.

Fourth, we describe an episome version of our BAC-TG EMBED system. In a single experiment, clones containing either stably integrated or extrachromosomally maintained BAC transgenes can be isolated. Most of the widely used episomal vectors are either based on viral sequences or derived from the non-viral pEPI plasmid (Van Craenenbroeck, Vanhoenacker and Haegeman, 2000; Conese, Auriche and Ascenzioni, 2004; Ehrhardt *et al.*, 2008; Lufino, Edser and Wade-Martins, 2008). A notable feature of the episomes generated by our BAC TG-EMBED system is that they are lost rapidly in the absence of drug selection, whereas both of the other two systems show selection-independent mechanisms for stable episome maintenance (Piechaczek *et al.*, 1999; Baiker *et al.*, 2000; Jenke *et al.*, 2004; Nanbo, Sugden and Sugden, 2007; Stehle *et al.*, 2007; Argyros *et al.*, 2008). While episome stability is valuable for certain applications, in other cases one would like to be able to easily eliminate the episomes as needed. Moreover, in contrast to the low copy number of episomes per cell produced using the other two methods, the BAC TG-EMBED method yields tens of BAC copies per cell allowing for much higher transgene expression levels. Additionally, the sizes of the episomes generated by the BAC TG-EMBED method are much larger than those generated by the other two methods. In the two clones we examined, the episomes were ~1 Mb and containing several copies of the BACs per episome and no detectable host DNA.

The high frequency creation and simple composition of these BAC episomes contrasts with human artificial chromosomes (HACs), which are special episomes, usually 1-10 Mb in size containing centromeric repeat sequences, mitotically stable, and

maintained at low copy number (Harrington *et al.*, 1997; Mills *et al.*, 1999; Kazuki and Oshimura, 2011; Kouprina *et al.*, 2013). Capable of introducing large DNA sequences into recipient cells, HACs have shown great potential in a wide range of applications, such as recombinant protein production, drug selection and gene therapy (Kazuki *et al.*, 2010; Takahashi *et al.*, 2010; Hiratsuka *et al.*, 2011; Kim *et al.*, 2011). However, the construction of HACs remains non-trivial: it requires cloning of either telomere sequences and/or alphoid DNA, the formation of HACs occurs at very low frequency and only in certain cell lines (Larin and Mejía, 2002), and the transfer of HACs from donor cells into recipient cells is difficult (Fournier and Ruddle, 1977; Liskovykh *et al.*, 2016). Moreover, the presence of large telomere sequences and/or alphoid DNA on the HACs, and the heterchromatic state associated with these repeats, increases the likelihood of transgene silencing.

In contrast, with our BAC-TG EMBED system, 10s-100s of stable cell clones containing multiple copies of ~Mb-size episomes, likely containing only BAC DNA, can be obtained from a single transfection. Cells containing high copy numbers of these BAC episomes can be enriched by flow sorting, while cells from these clones containing no BAC episomes can be recovered after removal of drug selection and/or flow sorting. We anticipate that with additional engineering, these BAC episomes might possibly become a high-capacity episome system complementary to HACs, assuming they can be isolated from one cell line and then introduced and propagated in other cell lines.

It remains unclear how these BAC episomes form in NIH 3T3 cells and why they do not do so in other cell lines. In the two clones we studied, the episomes were circular DNA and composed of several BAC copies. Interestingly, previous studies have shown

that plasmids containing a MAR that is also a replication initiation region (IR) could initiate gene amplification in certain primary cancer cells, forming homogenously staining regions (HSRs), integrating into existing double minutes (DMs) or forming DMs *de novo* in cells without DMs (Shimizu *et al.*, 2001, 2003). It is believed that the IR/MAR plasmids are initially replicated as extrachromosomal circles, and then they multimerize into larger circular molecules. These amplified circles further multimerize to form DMs, recombine with pre-existing DMs or integrate into chromosomes and initiate HSR formation (Shimizu, Shingaki and Kaneko-sasaguri, 2005; Shimizu, 2009). This model is very similar to the episome model of gene amplification, where instead of the IR/MAR plasmids, small extrachromosomal circular DNAs, which are several hundred kb in size and are possibly produced by small chromosome deletions, initiates DM and HSR formation (Carroll *et al.*, 1988; Schoenlein *et al.*, 1992; L'Abbate *et al.*, 2014).

Given that both the MAR and IR sequences are ubiquitous in the mammalian genome, it is likely that the BACs used in this study also contain MAR and/or IR sequences. However, unlike the MAR/IR plasmids, these BACs did not generate typical HSRs when integrated into the chromosomes, and the episomes were much smaller than DMs in NIH 3T3 cells. One possible explanation is that the BACs undergo initial steps of gene amplification to form the episomes in NIH 3T3 cells, but the cells have mechanisms to stop the episomes from further multimerization or amplification. As gene amplification happens only in cancer cells, perhaps BACs can only form episomes in certain cell lines. As shown here, BAC transgene formed episomes in a small fraction of HCT116 cells, which could not be stably maintained even with drug selection. Further study of this BAC episome phenomenon may provide new insights into the process of gene amplification.

Alternatively, the formation of BAC transgene episomes in NIH 3T3 cells might occur through a process completely unrelated to gene amplification. Future work will be needed to determine the actual mechanism of this BAC episomal formation in mouse 3T3 cells.

To facilitate the assembly of BACs expressing multiple mini-genes, we developed BAC-MAGIC, allowing creation of a multi-transgene expressing BAC in several weeks, rather than the 4-5 months which would have been required by multiple rounds of DNA insertion using conventional BAC recombineering. Initial attempts to reassemble large, ~50kb regions of DHFR using yeast DNA assembly failed, apparently due to recombination between repetitive elements within the DHFR BAC sequence as well as the expression cassettes. In contrast, assembly of 10-15 kb modules from several DNA fragments using yeast DNA assembly worked with high efficiency. BAC-MAGIC exploits Gibson and yeast DNA assembly to build smaller modules with efficient serial BAC recombineering to reconstruct large BAC constructs containing multiple mini-gene expression cassettes. More generally, BAC-MAGIC should provide a tool for reconstruction of large eukaryotic DNA sequences containing high numbers of repetitive elements.

Finally, as a demonstration of our new version of BAC TG-EMBED system, we created cell lines expressing three different fluorescently tagged proteins in a single stable transfection step requiring just several weeks to isolate and expand cell clones. Most cell clones expressed all three tagged proteins at uniform levels and at reproducible relative levels of expression. This contrasts with the 6-12 months we have devoted in previous studies to create similar cell lines expressing multiple tagged proteins (Khanna, Hu and

Belmont, 2014) through a series of individual transfections followed by extensive screening of colonies to identify the small fraction expressing suitable levels of tagged proteins with minimal variegation and/or progressive long-term transgene silencing over time.

We anticipate that our expanded BAC TG-EMBED toolkit similarly will facilitate a wide range of applications requiring simultaneous expression of multiple transgenes.

MATERIALS AND METHODS

PCR amplification of endogenous promoters

Primers (Data A.1) were designed using Primer3 (Rozen and Skaletsky, 2000) or NCBI primer blast (Ye *et al.*, 2012) to amplify 1-3 kb promoter regions which included either the entire or part of the 5' UTRs upstream of the first exons of target genes. We used human genomic DNA extracted from BJ-hTERT cells as the template for PCR. However, the UBC promoter, including a partially synthetic intron, was amplified from plasmid pUGG (Chaturvedi *et al.*, 2018).

Construction of dual reporter DHFR BACs

The original dual reporter BAC, DHFR-HB1-GN-HB2-RZ (Bian and Belmont, 2010), was derived from the CITB-057L22 BAC (DHFR BAC) containing mouse chr13:92992156-93161185 (mm9). DHFR-HB1-GN-HB2-RZ has an EGFP expression cassette inserted 26 kb downstream of the *Msh3* transcription start site, and a mRFP expression cassette inserted at 121 kb downstream of the *Msh3* transcription start site.

The EGFP expression cassette contains a CMV promoter-driven EGFP gene and a SV40 promoter-driven Kanamycin/Neomycin resistance gene, while the mRFP expression cassette has a CMV promoter-driven mRFP gene and a SV40-driven Zeocin resistance gene. New dual reporter DHFR BACs were created using a similar strategy to that used to create DHFR-HB1-GN-HB2-RZ, except that new mRFP expression cassettes were used, where the CMV promoter was replaced with alternative, human endogenous promoters. The intermediate DHFR BAC containing only the EGFP expression cassette, DHFR-HB1-GN (Bian and Belmont, 2010), was used to insert these new mRFP expression cassettes using λ Red-mediated homologous recombination (Warming *et al.*, 2005; Khanna *et al.*, 2013).

Plasmid p[MOD-HB2-CRZ] (Bian and Belmont, 2010) contains a CMV driven mRFP and a SV40 driven Zeocin resistance gene, flanked by two ~500 bp regions homologous to the DHFR BAC target site. Plasmid p[MOD-HB2-RCS-Zeo] was created by replacing the CMV-mRFP fragment between NotI and NheI sites of p[MOD-HB2-CRZ] with a synthetic DNA fragment "RCS" containing multiple rare restriction sites (Table A.1). The mRFP fragment generated by digesting p[MOD-HB2-CRZ] with NheI was then inserted into the NheI site of p[MOD-HB2-RCS-Zeo], yielding plasmid p[MOD-HB2-RCS-RZ]. The PCR-amplified endogenous promoters were then inserted into the RCS, generating plasmids p[MOD-HB2-promoter name-RZ]. Promoter functionality was tested by transient transfection of NIH 3T3 cells with these plasmids.

To insert the new mRFP expression cassettes into the DHFR-HB1-GN BAC, one round of λ Red-mediated recombination, using Zeocin resistance as positive selection, was performed according to a published protocol (Khanna *et al.*, 2013). DNA fragments

containing the new mRFP expression cassettes with a given promoter with flanking homologous arms were excised from p[MOD-HB2-promoter name-RZ] plasmids by PmeI. SW102, a derivative strain of *Escherichia coli* (*E. coli*), was used for recombination. Recombinants were selected on low-salt LB plates containing 25 μg/ml Zeocin and 12.5 μg/ml Kanamycin at 32°C for ~20 hours. Recombinant colonies were screened by PCR amplification of sequences flanking the site of insertion (primers listed in Data A.1). The integrity of BAC constructs was verified by restriction enzyme fingerprinting, where observed band patterns on agarose gels were compared with predicted ones.

Construction of BACs containing the UBC-GFP-ZeoR cassette

Construction of pUGG containing the UBC-GFP-ZeoR-FRT-GalK-FRT cassette was described previously (Chaturvedi *et al.*, 2018). Human BACs RP11-138I1 (UBB BAC), CTD-2643I7 (HBB BAC), CTD-2207K13 (2207K13 BAC) and mouse BAC RP23-401D9 (ROSA BAC) were obtained from Thermo Fisher Scientific. Mouse BAC CITB-057L22 (DHFR BAC) was a gift from Edith Heard (Curie Institute, Paris, France).

The UBC-GFP-ZeoR reporter gene insertion positions (mm9 or hg19) are chr17:16,301,887-16,301,888 in the UBB BAC, chr6:113,043,332-113,043,333 in the ROSA BAC, chr13:93,099,101-93,099,102 in the DHFR BAC, chr1:79,224,725-79,224,726 in the 2207K13 BAC, and chr11:5,390,233-5,390,244 in the HBB BAC.

λ Red-mediated BAC recombineering (Warming *et al.*, 2005; Khanna *et al.*, 2013) using a *galK*-based dual-selection scheme was used to introduce the UBC-GFP-ZeoR reporter cassette onto the BACs according to published protocols (Khanna *et al.*,

2013). DNA fragments with homology ends for recombineering were prepared by PCR using primers (Data A.1) with 74-bp homology sequences plus 16-bp sequences (forward, 5'- acagcagagatccagt-3'; reverse, 5'-tgttggctagtgcgt-3') that amplify the UBC-GFP-ZeoR-FRT-GalK-FRT cassette from plasmid pUGG. *E. coli* strain SW105 was used for BAC recombineering. Recombinants containing the UBC-GFP-ZeoR-FRT-GalK-FRT cassette were selected for *galK* insertion at 32°C on minimal medium in which D-galactose was supplied as the only carbon source. Recombinant colonies were screened using PCR with BAC specific primers flanking the target regions (Data A.1). Subsequently, FLP recombinase-mediated removal of *galK* from selected recombinant clones was done by inducing actively growing SW105 cells with 0.1% (w/v) L-arabinose. Negative selection against *galK* used minimal medium containing 2-deoxy-galactose; deletion of *galK* in recombinants was again verified using BAC specific primers (Data A.1). The integrity of BAC constructs was verified by restriction enzyme fingerprinting.

The UBB, HBB, 2207K13, ROSA, DHFR BACs with the UBC-GFP-ZeoR reporter gene inserted were named UBB-UG, HBB-UG, 2207K13-UG, ROSA-UG and DHFR-UG, respectively.

Cell culture and establishment of BAC cell lines

Mouse NIH 3T3 fibroblasts (ATCC CRL-1658^{TM}) were grown in Dulbecco's modified Eagle medium (DMEM, with 4.5 g/l D-glucose, 4 mM L-glutamine, 1 mM sodium pyruvate and 3.7 g/l $NaHCO_3$) supplemented with 10% HyClone Bovine Growth Serum (GE Healthcare Life Sciences, Cat. # SH30541.03). Human HCT116 cells (ATCC

CCL-247™) were grown in McCoy's 5A medium supplemented with 10% Fetal Bovine Serum (Seradigm, Cat. # 1500-500H).

BAC DNA for transfection of mammalian cells was prepared with the QIAGEN Large Construct Kit (QIAGEN, Cat. # 12462) as per the manufacturer's instructions. All BACs except DHFR BAC derived BACs were linearized before transfection: 2207K13-UG BAC with SgrAI (New England Biolabs, Cat. # R0603S), HBB-UG BAC with NotI (New England Biolabs, Cat. # R3189S) and all other BACs with the PI-SceI (New England Biolabs, Cat. # R0696S). Lipofectamine 2000 (Thermo Fisher Scientific, Cat. # 11668019) was used to transfect the cells with the BACs according to the manufacturer's directions. The dual reporter DHFR BACs and the BACs containing the UBC-GFP-ZeoR reporter gene were transfected into NIH 3T3. The 2207K13-UG BAC was also transfected into HCT116. The DHFR BACs containing the Lac operator repeats were transfected into an NIH 3T3 cell clone 3T3_LG_C29 stably expressing the EGFP-dimer LacI-NLS fusion protein (EGFP-LacI) (Bian *et al.*, 2013). Mixed clonal populations of stable transformants were obtained after ~2 weeks of selection (75 µg/ml Zeocin and 500 µg/ml G418 for NIH 3T3 cells transfected with the dual reporter DHFR BACs; 75 µg/ml or 200 µg/ml Zeocin for NIH 3T3 or HCT116 cells, respectively, transfected with the BACs containing the UBC-GFP-ZeoR reporter gene; 75 µg/ml Zeocin and 200 µg/ml Hygromycin B for 3T3_LG_C29 transfected with the DHFR BACs); individual cell clones were obtained by serial dilution or colony picking using filter discs (Strukov and Belmont, 2008).

To analyze the stability of reporter gene expression in NIH 3T3 cells, individual cell clones were grown continuously with or without Zeocin (75 µg/ml) selection for 96

days. We used the following clones (Figure 4 and Figure A.1): DHFR-UG BAC- f1-7, f3-13, f3-15 (uniform), f1-6, f2-1, f2-3 (heterogeneous); ROSA-UG BAC- 2D6- 3C11, 3D7 (uniform), 2C12, 3A1 (heterogeneous); UBB-UG BAC- 1C2, 1F1, 1F12, 2F5, 2G4, 4D3, 5C1, 5C7 (uniform), 1A8, 1D5, 6H2 (heterogeneous); 2207K13-UG BAC- 3E3, 5C8, 5E1, 6B9, 6E12, 6F4, 7B2 (uniform), 1E3, 6A2, 6C10, 7B9 (heterogeneous).

Flow cytometry

For analysis of reporter gene expression, cells were grown to ~40%-80% confluence, trypsinized, and resuspended in growth media at ~0.5-1 million/ml. For analysis of the expression of mRFP and EGFP, or mRFP alone, cell suspensions were run on a BD FACS AriaII (BD Biosciences) or a BD LSR Fortessa (BD Biosciences), using the PE channel (561 nm laser and 582/15 nm bandpass filter) for mRFP, and the FITC channel (488 nm laser, 505 longpass dichroic mirror and 530/30 nm bandpass filter) for EGFP. For analysis of GFP expression alone, the cell suspensions were run on a BD FACS Canto II Flow Cytometry Analyzer (BD Biosciences), using the FITC/Alexa Fluor-488 channel (488nm laser, 502 longpass dichroic mirror and 530/30 bandpass filter). Rainbow fluorescent beads (Spherotech, Cat. # RFP-30-5A) were used as fluorescence intensity standards. Each sample was run for 1-2 min or until the number of events after gating reached 10-20 thousand.

For cell sorting, cells were resuspended at ~10 million/ml in growth media and run on a BD FACS AriaII for up to 30-40 minutes. Sorting windows are shown in the main and supplementary figures (Appendix A).

Estimation of relative promoter strength

The red and green fluorescence of the mixed-clonal populations stably transfected with the dual-reporter DHFR BACs was measured by flow cytometry. The mean florescence values of all gated cells were divided by the bead intensity values for normalization. The ratio of normalized mRFP to normalized EGFP was calculated as a measure of promoter strength (Equation 2.1). All promoter strengths were then normalized with the CMV promoter strength (comparing the CMV-driven mRFP to the CMV-driven EGFP expression) to calculate the relative promoter strength (Equation 2.2) using the CMV promoter as the reference.

$$\text{promoter strength} = \frac{median(\text{PE}_{\text{cells}})/median(\text{PE}_{\text{beads}})}{median(\text{FITC}_{\text{cells}})/median(\text{PE}_{\text{beads}})} \quad 2.1$$

$$\text{relative promoter strength} = \frac{\text{promoter strength}_x}{\text{promoter strength}_{\text{CMV}}} \quad 2.2$$

Genomic DNA extraction

Genomic DNA was isolated by phenol/chloroform extraction (Sambrook and Russell, 2006b). Cultured cells were harvested and washed with 1x Cell Culture Phosphate Buffered Saline (PBS, Corning, Cat. # 21040CV). Sorted cells were pelleted. Up to ~2 million cells were resuspended in 100 µl High-TE buffer (10 mM Tris-Cl, pH 8, 10 mM EDTA, 25-100 µg/ml RNase A (QIAGEN, Cat. # 19101)) and lysed by adding 2.5 µl 20% SDS. After incubation at 37°C for several hours, the lysate was digested by ~0.2 mg/ml Proteinase K (New England Biolabs, Cat. # P8102 or P8107S) at 55°C for ~1 day. 1 M Tris-Cl (pH 8.0), 5 M NaCl and nuclease free water were added to the lysate to bring up

the total volume to ~600 µl and final concentrations of Tris-Cl to ~0.1 M and NaCl to ~0.2 M. The lysate was then extracted once with an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1 mixture, Fisher Scientific, Cat. # BP1752I-400) and once with an equal volume of chloroform/isoamyl alcohol (24:1 mixture, MilliporeSigma, Cat. # C0549). DNA was precipitated by adding 2.5 volumes of 100% ethanol, washed with 70% ethanol and resuspended in EB (10mM Tris-Cl, pH 8.5).

Estimation of transgene copy number

BAC or plasmid transgene copy number within individual cell clones or sorted cells was measured by real-time quantitative PCR (qPCR), using purified genomic DNA, iTaq universal SYBR Green Supermix (Bio-Rad Laboratories, Cat. # 1725121) and a StepOnePlus (Applied Biosystems). Relative quantitation methods were used for copy number calculation. Primers used for qPCR are listed in Data A.1. Mouse genes *Sgk1* and *Hprt1* were used as endogenous controls, assuming four copies of each gene per cell in NIH 3T3. For Figure 3d and Figure 2.5c, a primer pair (Zeo-GFP2for/rev) that binds to the UBC-GFP-ZeoR region was used to estimate transgene copy number. For Table 2.2, Table 2.3 and Figure A.5, in addition to Zeo-GFP2for/rev, 4 primer pairs binding to the DHFR BAC or 6 primer pairs binding to the HBB BAC were used to estimate the copy number of DHFR-UG or HBB-UG BAC, respectively. The ΔC_T method (Equation 2.3 and 2.4) was used to estimate the copy numbers of the PCR amplification regions on the UBC-GFP-ZeoR reporter gene or on the HBB BAC, and $\Delta\Delta C_T$ method (Equation 2.5 and 2.6) was used to estimate the copy numbers of the PCR amplification regions on the DHFR BAC. When multiple primer pairs were used for a region, the mean copy number

of all PCR amplification regions was calculated as the copy number of that region. Equation 2.3 and 2.7 were used to calculate the fold increase of BAC copy numbers in H1 and H2 samples relative to L.

$$\Delta C_T = C_{T\,\text{test region}} - (C_{T\,Sgk1} + C_{T\,Hprt1})/2 \quad 2.3$$

$$\text{copy number}_{\Delta C_T} = 4 \times 1.95^{-\Delta C_T} \quad 2.4$$

$$\Delta\Delta C_T = \Delta C_{T\,\text{transgene clone}} - \Delta C_{T\,\text{NIH 3T3}} \quad 2.5$$

$$\text{copy number}_{\Delta\Delta C_T} = 4 \times 1.95^{-\Delta\Delta C_T} \quad 2.6$$

$$\text{BAC fold increase} = 1.95^{\Delta C_{T\,\text{L}} - \Delta C_{T\,\text{H1|H2}}} \quad 2.7$$

Correlation of reporter gene expression and reporter gene copy number

Mean fluorescence intensity (in arbitrary units) of individual clones were measured by flow cytometry and normalized by fluorescent bead intensity to be used as a measure of reporter gene expression. To ensure uniform normalization for all samples, fluorescent beads from the same batch were used for all measurements. Untransfected cells were used to establish background fluorescence levels. Linear correlations of GFP expression level versus transgene copy number for each group of cell clones were calculated using the linear trend line tool in Microsoft Excel with the y-intercept fixed to 0 (autofluorescence normalized by beads was almost 0).

DNA FISH probes

Biotin or digoxigenin labeled DNA FISH probes were made from BAC DNA, using a published protocol (Dernburg, 2011), with the following reagents: AluI, DpnI,

HaeIII, MseI, MspI, RsaI (New England Biolabs, Cat. # R0137S, R0176S, R0108S, R0525S, R0106S, R0167S, respectively) and CutSmart Buffer (New England Biolabs); Terminal Deoxynucleotidyl Transferase and reaction buffer (Thermo Fisher Scientific, Cat. # EP0161); dATP (New England Biolabs, Cat. # N0446S) and Biotin-14-dATP (Thermo Fisher Scientific, Cat. # 19524016) for biotin labelling, or dTTP (New England Biolabs, Cat. # N0446S) and Digoxigenin-11-dUTP (MilliporeSigma, Cat. # 11093088910) for digoxigenin labelling.

3D DNA FISH

DNA FISH of interphase nuclei used published protocols (Cremer *et al.*, 2008; Solovei and Cremer, 2010) with small modifications. Cells grown on coverslips (12 mm diameter) were fixed with 3-4% paraformaldehyde in Dulbecco's phosphate buffered saline (DPBS, 8 g/l NaCl, 0.2 g/l KCl, 2.16 g/l Na_2HPO_4-$7H_2O$, 0.2 g/l KH_2PO_4) for 10 min, followed by permeabilization with 0.5% Triton X-100 (Thermo Fisher Scientific, Cat. # 28314) in DPBS for 10-15 min. Cells were subjected to six freeze-thaw cycles using liquid nitrogen, immersed in 0.1M HCl for 10-15 min, and then washed 3x with 2x saline-sodium citrate (SSC). Freeze-thaw cycles sometimes were skipped with no noticeable difference in FISH signals. Cells were incubated in 50% deionized formamide (MilliporeSigma, Cat. # S4117)/2x SSC for 30 min at room temperature (RT), and stored for up to 1 month at 4°C. Each coverslip used ~4 µl hybridization mixture, consisted of 5-20 ng/µl probes, 10x of mouse (for NIH 3T3 cells) or human (for HCT116 cells) Cot-1 DNA (Thermo Fisher Scientific Cat. # 18440016 or 15279011,) per ng probe, 50% deionized formamide, 10% dextran sulfate (MilliporeSigma, Cat. # D8906) and 2x SSC.

Cells and probes were denatured together on a heat block at ~76°C for 2-3 min and hybridized at 37°C for 16 hrs-3 days. After hybridization, cells were washed 3 x 5 min in 2x SSC at RT, and for 3 x 5 min in 0.1x SSC at 60°C, and then rinsed with SSCT (4x SSC with 0.2% TWEEN 20) at RT. FISH signals were detected by incubation with Alexa Fluor 647 conjugated Strepavidin (1:200; Jackson ImmunoResearch, Cat. # 016-600-084) or Alexa 594 conjugated Strepavidin (1:200; Life Technology, Cat. # S11227) for biotin-labeled probes, or Alexa Fluor 647 conjugated IgG fraction monoclonal mouse anti-digoxin (1:200; Jackson ImmunoResearch, Cat. # 200-602-156) for digoxigenin labeled probes, diluted in SSCT with 1% Bovine Serum Albumin (MilliporeSigma, Cat. # A7906), for 40 min-2 hrs at RT. Coverslips were washed in SSCT for 4 × 5 min, rinsed with 4x SSC and mounted.

Mitotic FISH

Metaphase spreads were prepared according to a published protocol (Beatty and Scherer, 2002) with small modifications. Cells grown to 70-80% confluence were incubated with 0.1 μg/ml Colcemid (Thermo Fisher Scientific, Cat. # 15212012) in growth media for ~1 hr. Cells were then harvested and swollen by incubation in 0.075 M KCl for 10-20 min at 37°C, followed by fixation with freshly prepared Carnoy's fixative (3:1 v/v ratio of methanol/acetic acid). Chromosomal spreads were made by dropping the fixed swollen cells onto cold wet glass slides. DNA FISH of mitotic spreads was performed using a published protocol (Beatty and Scherer, 2002).

Microscopy and image analysis

For examining EGFP-LacI signals cells were grown on coverslips and fixed with 3-4% paraformaldehyde in DPBS before mounting. For examining the expression of the three reporter minigenes, SNAP tagged-Lamin B1, SNAP-tagged Fibrillarin and mCherry-Magoh, the cells were first labeled with cell-permeable substrate SNAP-Cell Fluorescein (New England Biolabs, Cat. # S9107S) overnight at 240 nM concentrations. To reduce background of unreacted SNAP-tag substrate, cells were incubated 3x 30 mins with media in the incubator, washed with PBS, and fixed with freshly prepared 4% paraformaldehyde in PBS for 15 min at RT. All samples- including fixed cells expressing fluorescently tagged transgenes, 3D DNA FISH, and mitotic FISH sample- were mounted with a Mowiol-DABCO anti-fade medium ('Mowiol mounting medium', 2006) containing ~3 μg/ml DAPI (MilliporeSigma, Cat. # D9542).

3D z-stack images were acquired using a Deltavision wide-field microscope (GE Healthcare), equipped with a Xenon lamp, 60X, 1.4 NA oil immersion objective (Olympus) and CoolSNAP HQ CCD camera (Roper Scientific) or a V4 OMX (GE healthcare) microscope, equipped with a 100X, 1.4 NA oil immersion objective (Olympus) and two Evolve EMCCDs (Photometrics). Images were deconvolved using the deconvolution algorithm ('Mowiol mounting medium', 2006) provided by the *softWoRx* software (GE Healthcare). Gamma = 0.5 was applied to green channels in Figure 2.5e, Figure A.8 and Figure A.10 for proper display of spots with relatively low signals. All image analysis and preparation were done using Fiji (Schindelin *et al.*, 2012). Images were assembled using Illustrator (Adobe), Photoshop (Adobe), or GIMP.

For estimation of episome size, the z-sections containing focused episome images for the DAPI and FISH channels were selected manually from the deconvolved z-stack image. Chromosomes and FISH spots were segmented by applying the k-mean clustering algorithm (number of clusters = 3, cluster center tolerance = 0.0001, randomization seed = 48) from the IJ Plugins Toolkit (http://ij-plugins.sourceforge.net/plugins/toolkit.html). The smallest chromosome was identified by manually searching for the chromosome with the smallest area. Segmented FISH spots overlapping or touching chromosomes were removed manually. Integrated DAPI intensities of the smallest chromosome and of the FISH spots not overlapping or touching chromosomes were calculated by Equation 2.8 (Mean gray value and Area were measured by Fiji). Average episome size was calculated by Equation 2.9 (n is the number of FISH spots, chro is the smallest chromosome found in the field, 61.4 Mb is the size of chr19 in mm10).

$$\text{integrated density} = (\text{Mean gray value} - 200) \times \text{Area} \tag{2.8}$$

$$\text{episome size} = \frac{\text{integrated density}_{\text{FISH}}}{n \times \text{integrated density}_{\text{chro}}} \times 61.4 \text{ Mb} \tag{2.9}$$

Comparison of reporter gene expression levels in for NIH 3T3 cell clones (Figure 2.7c) was done by projecting deconvolved images stacks and then measuring the integrated intensity within individual nuclei after subtracting background intensity levels measured in the cytoplasm. Regions of interest circumscribing individual nuclei were drawn manually based on the SNAP-lamin B1 signal. Linear correlations of the integrated intensities of the nuclear SNAP-tag and mCherry signals were calculated using Microsoft Excel with the y-intercept fixed to 0.

A non-linear Gamma correction (0.7) to reduce the grey-scale dynamic range followed by a maximum intensity projection of 3-4 z-sections was used to better visualize both lamin and nucleolar staining simultaneously (Figure 2.7d).

Agarose embedded DNA preparation and S1 Nuclease digestion

Agarose embedded DNA was prepared according to published protocols (Sambrook and Russell, 2006a; Khan and Kuzminov, 2017) with modifications. To prepare mammalian cell suspensions, cells were grown without selection for 3-4 days after passaging, reaching 80%-90% confluence. Cells were trypsinized, resuspended in cell media, washed with PBS, and resuspended in PBS at a concentration of ~8 x 10^6 cells / 100 µl. To prepare *E. coli* cell suspensions, ~0.1 ml of overnight culture was diluted in 15 ml fresh LB and grown to an OD_{600} of ~1. Cells were washed with L Buffer (10 mM Tris-Cl pH7.6, 20 mM NaCl, 100 mM EDTA) once and resuspended in L Buffer at a concentration of ~10^9/100 µl, assuming a cell concentration of ~8 x 10^7/100 µl at an OD_{600} of 1.

2% certified low melt agarose (Bio-Rad Laboratories, Cat. # 1613111) was prepared with L Buffer and kept at 75°C. Equal volumes of the cell suspension (RT) and the agarose solution (75°C) were mixed and immediately transferred to plug molds (Bio-Rad Laboratories, Cat. # 1703713), ~100 µl mixture per plug. The agarose plugs were incubated in L Buffer with 1% Sarcosyl (MilliporeSigma, Cat. # L5125) and 0.5 mg/ml proteinase K at 55°C for 1-2 days. The agarose plugs were washed with W Buffer (20 mM Tris-Cl, pH7.6, 50 mM EDTA) for 2 x 15 min, incubated in 1 mM PMSF in W

Buffer for 30 min, and washed with W Buffer again. Prepared agarose plugs were stored in 0.5 M EDTA at 4°C before use.

For S1 Nuclease (Promega, Cat. # M5761) digestion, agarose plugs were first washed in TE (10 mM Tris-Cl, 1 mM EDTA, pH 7.6) for 3 x 10 min and in 1x S1 Nuclease Buffer for 20 min on ice. The agarose plugs were then digested with 1-16 U/0.4 ml S1 Nuclease in 1x S1 Nuclease Buffer at 37°C for 45 min. The reaction was stopped by washing the agarose plugs with 0.5 M EDTA or W Buffer.

Pulsed Field Gel Electrophoresis (PFGE)

PFGE was performed using a CHEF-DR III (Bio-Rad Laboratories) according to the manufacturer's manual using a 1% certified megabase agarose (Bio-Rad Laboratories, Cat. # 1613108) gel in 0.5x Tris-borate-EDTA buffer (TBE), a 0.5x TBE running buffer, and the following parameters: voltage = 6 V/cm, angle = 120°, pulse = 60-120 sec, temperature = 14 °C, run time = 20 or 24 hrs (stopped at 18-20 hrs). Yeast chromosomes (Bio-Rad Laboratories, Cat. # 170-3605) were used as DNA size markers.

Southern hybridization probes

Southern hybridization probes were created and labeled with digoxigenin by PCR using primers listed in Data A.1. Set 1 contains a 620bp and a 615 bp fragment amplified from the GFP-ZeoR region; Set 2 contains 525 bp, 534 bp, and 504bp fragments amplified from the BAC vector region; Set 3 contains 446 bp, 681 bp, and 424 bp fragments amplified from the HBB BAC. Pooled Set 1 and Set 2 fragments were used for detecting the DHFR BAC, and pooled Set 1 and Set 3 for detecting the HBB BAC. PCR

was done using *Taq* DNA polymerase (New England Biolabs, Cat. # M0267L) with the following recipe: 1x ThermoPol Buffer, 0.2 mM dATP/dCTP/dGTP (New England Biolabs, Cat. # N0446S), 0.165 mM dTTP (New England Biolabs, Cat. # N0446S), 0.035 mM Digoxigenin-11-dUTP, 0.5 ng HBB BAC, 1.25 U *Taq* DNA polymerase, 0.5 μM forward/reverse primers, 50 μl total reaction volume. PCR products were column (QIAGEN, Cat. # 28104) purified. Pooled probes were denatured in nuclease free water, at ~100°C for ~10 min and snap-chilled on ice before use.

Southern hybridization

Southern blotting used a published protocol (Kimura *et al.*, 2010) with modifications. After ethidium bromide staining and imaging, the gel was depurinated in 0.25 M HCl for 2x 30 min, denatured in 0.4 M NaOH for 2x 25 min, neutralized in 0.5 M Tris-Cl/1.5 M NaCl (pH 7.6) for 2x 20 min and washed in 2x SSC for 2x 20 min. DNA was transferred to Zeta-Probe membranes (Bio-Rad Laboratories, Cat. # 1620165) using a Model 785 Vacuum Blotter (Bio-Rad Laboratories), with 2x SSC as transfer buffer, ~5 inches Hg pressure, and ~16 hrs transfer time. A Stratalinker (Strategene) was used to cross-link DNA to the membrane.

Hybridization used a standard protocol (Sambrook and Russell, 2006c) with modifications. The hybridization buffer was composed of 1:1 volumes of 1 M Na_2HPO_4 (pH 7.2) and 14% (w/v) SDS. Total concentration of pooled probes was ~100 ng/ml. Hybridization was carried out at 65°C for ~16 hrs. After hybridization, the membrane was washed with 2x SSC/0.1% SDS for 2 x 5 min at room temperature, and with 1x SSC/0.1% SDS for 2 x 10 min at 65°C and rinsed with 2x SSC. Signals were detected

using the DIG Nucleic Acid Detection Kit (MilliporeSigma, Cat. # 000000011175041910) according to the manufacturer' manual, except that in the final step, CDP-*Star* (MilliporeSigma, Cat. # 11685627001) was used instead of NBT/BCIP, and the membrane was imaged by an iBright system (Thermo Fisher Scientific).

Estimation of average BAC DNA content per episome

To estimate the average BAC DNA content per episome of clone DHFR-UG-s3 and clone HBB-UG-100d3, cells at the same passage were seeded on glass coverslips for DNA FISH using BAC probes, and in different plates for genomic DNA extraction followed by qPCR. The mean number of FISH spots per nucleus, counted from z-stack projected images, provided the average episome copy number per cell. For the DHFR-UG clone, 3 was subtracted from the mean number of FISH spots, as the parental NIH 3T3 cells had ~3 FISH spots, corresponding to the endogenous DHFR loci, using FISH probes prepared from the DHFR BAC. qPCR estimation of BAC copy number per cell was described in section "Estimation of transgene copy number". BAC DNA content per episome was calculated using Equation 2.10.

$$\text{BAC content per episome} = \frac{\text{BAC copy number per cell}}{\text{episome copy number per cell}} \times \text{BAC size} \quad 2.10$$

Whole genome sequencing

Clone DHFR-UG-s3 and clone HBB-UG-100d3 were sorted by flow cytometry using the H1, H2 and L sorting windows shown in Figure 2.6c and Figure A.4a. Cells from the H2 and L regions were sorted in the same experiment, while cells from the H1

regions were sorted in another experiment. 100-200 thousand cells were collected from each window. Genomic DNA from sorted cells was isolated by phenol-chloroform extraction. To prepare sequencing libraries, genomic DNA was first fragmented to 100-500 bp by sonication using a Bioruptor Pico (Diagenode), with the following conditions: 4 ng/μl DNA in 120 μl EB, 1.5 ml tube, 10-11 cycles of 30 secs on and 30 secs off. Next, indexed adaptors was attached to the fragmented DNA using True-Seq ChIP Sample Preparation kit (Illumina, Cat. # IP-202-1012) according to the manufacturer's instructions with the following modifications: after the fragmented DNA was end repaired, 3' end adenylated, and ligated to indexed adaptors without size selection, the ligation products were PCR amplified for 7~9 cycles. Libraries were quality checked on a Fragment Analyzer (Agilent) and quantitated by qPCR. Every 6 libraries were pooled at equal molar ratios and sequenced on one lane using a HiSeq 4000 for 101 cycles from one end of the fragments using a HiSeq 4000 sequencing kit version 1. Fastq files were generated and de-multiplexed with the bcl2fastq v2.20 Conversion Software (Illumina). Library quality checking, quantitation and sequencing, and fastq file generation and de-multiplexing were done by the DNA services lab, Roy J. Carver Biotechnology Center, UIUC. 59-65 million reads with quality score >30 were obtained for each library.

Sequencing reads processing and copy number variation analysis

Low quality bases and adaptor sequences were trimmed from raw reads using cutadapt 1.14 with Python 2.7.13 with the following parameters: -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -q 20,20 -m 20, resulting in ~0.2% bp being trimmed. Reads were then aligned to a reference genome (mm10 plus

HBB BAC (CTD-2643I7, sequence from hg38), the BAC vector (pBelo11, GenBank Accession #: U51113) and UBC-GFP-ZeoR, each as an individual chromosome) using Bowtie2 (version 2.3.2) with default parameters. Overall alignment rate of each sample was ~98%-99%. Finally, PCR duplicates were removed by SAMtools rmdup (version 1.7) with default parameters, resulting in 42-48 million total mapped reads in each sample.

For reads binning, each chromosome of the reference genome was divided into non-overlapping 3 kb or 30 kb bins; the number of alignments with centers falling into each bin (binned reads) was counted and then divided by the mean read count (Equation 2.11), generating normalized binned reads (normalized reads, Equation 2.12), and finally the normalized binned reads of the test sample (H1 or H2 cells) were divided by that of the reference sample (L cells), generating the ratio of normalized binned reads (ratio, Equation 2.13). The mean read count was ~50 or ~500 for 3 kb or 30 kb bin size, respectively. To reduce noise caused by extremely low read counts, a threshold for determining outliers was calculated based on the quantile range (Equation 2.14). Bins with $\log_2$(reads) smaller than the threshold in the test sample were removed from further analysis. The excluded bins took up ~6.0% of total bins for both 3 kb and 30 kb bin sizes, including zero read count bins, which took up ~5.5% or ~3.5% of total bins for 3 kb or 30 kb bin size, respectively. The maximum number of reads of the excluded bins were ~7 or ~108 for 3 kb or 30 kb bin size, respectively.

$$\text{mean read count} = \frac{\text{total mapped reads} \times \text{bin size}}{\text{reference genome size}} \qquad 2.11$$

$$\log_2(\text{normalized reads}) = \begin{cases} \log_2 \dfrac{\text{binned reads}}{\text{mean read count}}, & \text{binned reads} > 0 \\ \log_2 \dfrac{0.1}{\text{mean read count}}, & \text{binned reads} = 0 \end{cases} \quad 2.12$$

$$\log_2(\text{ratio}) = \log_2 \frac{\text{reads}_{\text{H1|H2}}}{\text{reads}_{\text{L}}} \quad 2.13$$

$$\text{outlier threshold} = 25\%\ \text{quantile} - 4 \times (75\%\ \text{quantile} - 25\%\ \text{quantile}) \quad 2.14$$

A circular binary segmentation algorithm (Olshen *et al.*, 2004; Venkatraman and Olshen, 2007) from the R-package DNAcopy (version 1.52.0) was used to merge bins with similar $\log_2$(ratio) into segments, with the following parameters for the segment function: verbose = 1, undo.splits="sdundo", undo.SD=1. The mean $\log_2$(ratio) of each segment was calculated for identifying episome-localizing regions.

To identify possible episome-localizing regions, we first measured BAC transgene copy numbers in the H1, H2 and L samples by qPCR and then calculated the theoretical episome copy numbers using the estimated BAC copy number per episome of unsorted cells (Table 2.2). The minimum copy number increase of episome-localizing host DNA (minimum increase) was then calculated assuming NIH 3T3 to be tetraploid and each episome to have the same host DNA sequence (Equation 2.15). Segments with mean $\log_2$(ratio) equal to or greater than $\log_2$(minimum increase) in both H1 and H2 samples were selected as candidate episome-localizing regions.

$$\text{minimum increase}_{\text{H1|H2}} = \frac{\text{episome copy number}_{\text{H1|H2}} + 4}{\text{episome copy number}_{\text{L}} + 4} \quad 2.15$$

Construction of DHFR BAC deletions

We tested several DHFR BAC deletions- made for other purposes- for their ability to produce episomes. The DHFR-c27 BAC (Bian and Belmont, 2010) containing a 256-mer Lac operator (LacO) repeats and a CMV-mRFP-SV40-ZeoR expression cassette was derived from the DHFR BAC, and was used for making the DHFR BAC deletions. DHFR-c27d2 contains a ~70 kb deletion of the 3' part of the *Msh3* gene. DHFR-c27d3-crz contains a ~80 kb deletion of the whole *Dhfr* gene and the 5' part of the *Msh3* gene, including the CMV-mRFP-SV40-ZeoR expression cassette inserted in the *Msh3* gene, and contains a new CMV-mRFP-SV40-ZeoR expression cassette introduced at the remaining part of *Msh3* gene. DHFR-c27d4 contains a ~20 kb deletion around the divergent promoter region. λ Red-mediated BAC recombineering with a *galK*-based dual-selection scheme was used to create the deletion BACs from the DHFR-c27 BAC, as described in "Construction of dual reporter DHFR BACs" and "Construction of BACs containing the UBC-GFP-ZeoR cassette". DNA fragments containing either GalK or FRT-GalK-FRT and homology ends were produced by PCR using either pGalK or pUGG as templates. For DHFR-c27d2, the GalK cassette was introduced by the first round of recombination and was subsequently removed by another round of recombination using a DNA fragment created by a pair of partially overlapping primers. For DHFR-c27d3 and DHFR-c27d4, the FRT-GalK-FRT cassette was introduced in the first round instead and was subsequently removed by inducing FLP recombinase as described in "Construction of BACs containing the UBC-GFP-ZeoR cassette". To create DHFR-c27d3crz, the CMV-mRFP-SV40-ZeoR cassette was introduced into DHFR-c27d3 by one round of recombination using Zeocin resistance as positive selection as described in "Construction of dual reporter DHFR BACs". Each

round of recombination was validated by PCR and restriction enzyme fingerprinting. All primers are listed in Data A.1.

Construction of multi-reporter DHFR BAC by BAC-MAGIC

Overview: Construction of the 3-reporter BAC was done by serially inserting ~10-15 kb DNA cassettes into the DHFR BAC scaffold by BAC recombineering. These DNA cassettes were constructed from two different DNA plasmid module types: reporter modules and intervening DHFR sequence modules. DNA cassettes were inserted sequentially into the DHFR BAC using multiple rounds of BAC recombineering and positive selection with one of two different positive selectable markers. After insertion of the first DNA cassette, each subsequent insertion of the next DNA cassette removed the preceding positive selectable located at the 3' end of the preceding cassette while inserting the alternative selectable marker located at the 3' end of the new cassette. Three reporter gene modules (Rep Mod 01, 02, 03) plus three intervening DHFR sequence modules (DHFR 02, 03, 04) were constructed and then inserted into the DHFR BAC using 6 sequential rounds of BAC recombineering. In this way, 45 kb of the original DHFR BAC effectively was reconstructed such that the original DHFR sequences were retained but the 3 reporter mini-genes were inserted into this BAC region with each reporter minigene spaced by ~10 kb of DHFR sequence. We call this overall construction approach BAC-MAGIC (**BAC**-**M**odular **A**ssembly of **G**enomic loci **I**nterspersed **C**assettes).

Each DNA cassette was constructed using traditional cloning methods, Gibson assembly (Gibson *et al.*, 2009), and/or DNA Assembler (Shao, Zhao and Zhao, 2009;

Shao and Zhao, 2012). Three reporter recipient modules (pRM01-Spec, pRM02-Spec, and pRM03-Spec) were designed to incorporate a rare AgeI restriction site for insertion of reporter expression cassettes of choice, in order to create the final reporter modules for BAC recombineering. Unless mentioned specifically all the enzymes were procured from New England Biolabs. All primers and oligos are listed in Data A.1. Gibson assembly used Gibson assembly cloning kit (New England Biolabs, Cat. # E5510S) as per the manufacturer's instructions.

DNA Assembler used *Saccharomyces cerevisiae (S. cerevisiae)* strain VL6-48N (MATα, his3-Δ200, trp1-Δ1, ura3-Δ1, lys2, ade2-101, met14, cir°), transformed with 43 fmol pRS413 vector backbone and 130 fmol of all other fragments using the LiAc/SS carrier DNA/PEG method (Gietz and Schiestl, 2007). The *S. cerevisiae* single-copy shuttle vector pRS413 contains CEN6/ARS autonomously replicating sequence, auxotrophic selection marker *HIS3* for propagation in yeast, and pMB1 origin of replication and *bla* (Ap^R) marker for selection with ampicillin in *E. coli*. The 3.8 kb pRS413 vector backbone was PCR amplified from plasmid pRS413 (New England Biolabs) using primer pair RS413-Fw/RS413-Rev for all yeast assembly reactions. The vector backbone and all other fragments made by PCR were digested with DpnI to remove template DNA. Transformants were selected on SC selection media plates lacking histidine [0.17% Bacto-yeast nitrogen base without amino acids (MilliporeSigma, Cat. # Y1251-100G), 0.5% ammonium sulfate, 2% D-glucose, 0.2% Dropout mix (MilliporeSigma, Cat. # Y2001-20G), 2% agar, 80 mg/l uracil, 80 mg/l L-tryptophan, and 240 mg/l L-leucine] at 30°C for 3-4 days. Plasmid DNA were prepared using QIAprep Spin Miniprep Kit (Qiagen, Cat. # 27104) and screened by restriction enzyme

fingerprinting. Plasmid DNA from selected yeast colonies was introduced into *E. coli* strain DH5α and isolated plasmid DNA then further validated by additional restriction enzyme fingerprinting.

Below we describe construction of each reporter and intervening spacer modules and BAC recombineering assembly of these modules to create the 3-reporter BAC.

Construction of plasmid pRM01-RSLB1-Spec (Reporter module 01): Plasmid pRM01 was made by sequential addition of two DHFR homology regions to plasmid pEGFP-C1 (Clontech). First, the 2.1 kb DHFR homology region (M1F4) was PCR amplified from the DHFR BAC using primer pair M1F4-BamHIfor/M1F4-AgeIrev, double digested with BamHI/AgeI, and ligated with the BamHI/AgeI digested pEGFP-C1 to generate intermediate plasmid pEG-Rep-Module-1a. Next, the 2.0 kb DHFR homology region (M2F12) was PCR amplified from the DHFR BAC using primer pair M2F12-AgeIFor/M2F12-PshRev, double digested with AgeI/PshAI and ligated with the AgeI/SnaBI digested plasmid pEG-Rep-Module-1a to produce plasmid pRM01.

To create plasmid pRM01-Spec (Reporter recipient module 01), a 1.6 kb Spectinomycin resistance gene expression cassette (SpecR), derived from plasmid pYES1L (Thermo Fisher Scientific), was inserted into pRM01, 400 bp upstream of the 3' end of the M2F12 DHFR homology region by two-fragment Gibson Assembly (Gibson *et al.*, 2009). The two fragments for Gibson assembly were PCR amplified from pRM01 using primer pair GA-RM01-Spec-For/ GA-RM01-Spec-Rev (PCR product size: 7.8 kb), or from pYES1L using primer pair Specfor/SpecRev (PCR product size: 1.6 kb) respectively.

The pRSLB1 (hRPL32-SNAP-Lamin B1) plasmid harboring SNAP-tagged Lamin B1 reporter expression cassette (RSLB1) was constructed by three-fragment Gibson assembly. pEGFP-Lamin B1 plasmid vector backbone 5.3 kb fragment was prepared by AseI/BsrGI double digestion. The hRPL32 promoter (2.2 kb) and SNAP tag (561 bp) fragments were PCR amplified using primer pairs GA-hRPL32-fwd/GA-hRPL32-rev (template plasmid pMOD-HB2-hRPL32-RZ, made in this study), and GA-SNAP-fwd/GA-SNAP-rev (template plasmid pSNAPf, New England Biolabs).

pRM01-Spec was linearized by AgeI and simultaneously dephosphorylated by Shrimp Alkaline Phosphatase (New England Biolabs, Cat. # M0371S). The RSLB1 expression cassette was PCR amplified from plasmid pRSLB1 using primer pair R32CerLBAgeIfor/newPCFAgeIrev (PCR product size: 4.9 kb) and double digested with DpnI/AgeI. The linearized pRM01-Spec and the digested RSLB1 PCR product were ligated to produce plasmid pRM01-RSLB1-Spec, which was digested with AseI to produce the final BAC recombineering 10.3 kb targeting construct.

Construction of plasmid pRM02-PSF-Spec (Reporter module 02): Plasmid pRM02 was made using similar cloning steps used to produce pRM01 except two different DHFR homology regions were added to pEGFP-C1: 2.0 kb PCR product M2F4 (primer pair M2F4-BamHIfor/M2F4-AgeIrev) replaced M1F4 and 2.0 kb PCR product M3F1 (primer pair M3F1-AgeIFor/M3F1-PshRev) replaced M2F12. Plasmid pRM02-Spec was made the same way as pRM01-Spec except that fragment 1 for Gibson assembly was PCR amplified from plasmid pRM02 using primer pair GA-RM02-Spec-For/GA-RM02-Spec-Rev (PCR product size: 7.8 kb). The final plasmid pRM02-Spec

(pRep-module 02-Spec) is Reporter recipient module 02 for the SNAP-tagged Fibrillarin reporter expression cassette (PSF).

To create plasmid pPSF (pPPIA-SNAP-Fibrillarin), the GFP cassette between KpnI/HpaI restriction sites of plasmid GFP-Fibrillarin was replaced with a 730 bp Cerulean cassette PCR amplified from plasmid pCerulean-N1 (New England Biolabs) using primer pair ForCerFib/RevCerFib, resulting in an intermediate plasmid pPCF. Next, the CMV promoter between SnaBI/HindIII sites of pPCF was replaced with the 2.8 kb PPIA promoter PCR amplified from plasmid p[MOD-HB2-PPIA-RZ] (made in this study) using primer pair PPIACerFibFor/ PPIACerFibRev, resulting in plasmid pPPIA-Cer-Fib. Finally, the 720 bp Cerulean cassette between the AgeI/HpaI sites of pPPIA-Cer-Fib was replaced with a 560 bp SNAP tag fragment PCR amplified from plasmid pSNAPf (New England Biolabs) using primer pair Snap-XmaI-For/Snap-HpaI-Fib-Rev and double digested with XmaI/HpaI, producing pPSF.

pRM02-Spec was linearized by AgeI and simultaneously dephosphorylated by Shrimp Alkaline Phosphatase (New England Biolabs, Cat. # M0371S). The 4.6 kb PSF expression cassette was PCR amplified from pPSF using primer pair PSF-AgeI-For/ PSF-AgeI-Rev and double digested with DpnI/AgeI. Their ligation produced plasmid pRM02-PSF-Spec, which provided the 10.4 kb BAC recombineering targeting construct after BamHI/AatII/RsrII triple digestion of pRM02-PSF-Spec.

Construction of plasmid pRM03-PCM-Spec (Reporter module 03): Plasmid pRM03 was made using similar cloning steps used to produce pRM01 except two different DHFR homology regions were added to pEGFP-C1: 2.1 kb PCR fragment M3F4 (primer pair M3F4-BamHIfor/M3F4-AgeIrev) replaced M1F4 and 2.1 PCR

fragment M4F1 (primer pair M4F1-AgeIFor/M4F1-PshRev) replaced M2F12. Plasmid pRM03-Spec was made the same way as pRM01-Spec except that fragment 1 for Gibson assembly was PCR amplified from plasmid pRM03 using using primer pair GA-RM03-Spec-For/ GA-RM03-Spec-Rev (PCR product size: 7.8 kb). The final plasmid pRM03-Spec (pRep-module 03-Spec) is Reporter recipient module 03 for the mCherry-tagged Magoh reporter expression cassette (PCM).

Plasmid pPCM (pPPIA-mCherry-Magoh) was created in two steps. First, the CMV promoter between the NdeI/NheI sites of plasmid pmRFP-Magoh was replaced with the PPIA promoter (2.8 kb), PCR amplified from plasmid pMOD-HB2-PPIA-RZ using primer pair PPIA-Magohfor/ PPIA-MagohRev and double digested with NdeI/NheI, resulting in intermediate plasmid pPMM. Next, the mRFP tag between the NheI/HindIII sites of pPMM was replaced with a 720 bp mCherry tag PCR amplified from plasmid pQCXIN-TetR-mCherry using primer pair mCherry-NheI-Magoh-For/mCherry-H3-Magoh-Rev, resulting in plasmid pPCM.

To create plasmid pRM03-PCM-Spec (Reporter module 03), plasmid pRM03-Spec was linearized by AgeI and simultaneously dephosphorylated by Shrimp Alkaline Phosphatase (New England Biolabs, Cat. # M0371S). A 4.2 kb PCM expression cassette was PCR amplified form plasmid pPCM using primer pair MMorCF-AgeIfor/newPCFAgeIrev and double digested with DpnI/AgeI. pRM03-Spec and the PCM PCR product were ligated, producing plasmid pRM03-PCM-Spec, which was used as a template for PCR amplification using primer pair M3F4-PCR-Fw/M4F1-PCR-Rev to produce the 9.9 kb BAC recombineering target. After PCR, any remaining template plasmid was digested with DpnI.

Construction of plasmid pRS413-DHFR-Mod-02-Kan (Intervening DHFR module 02): Plasmid pRS413-DHFR-Mod-02 was made by assembling the vector backbone with four additional fragments using the DNA assembler method (Shao, Zhao and Zhao, 2009; Shao and Zhao, 2012). Fragment 5'-DHM2 (4.3 kb) and fragment 3'-DHM2 (6.3 kb) with an overlap of 659 bp and were both PCR amplified from the DHFR BAC, using primer pair M2F12-AgeIfor/M2F1rev or DHM2-Seq2/M2F4-AgeIrev, respectively. Two bridging oligomers, with a 125 bp homology to the pRS413 vector backbone, and a 125 bp homology to fragment 5'-DHM2 (oligo M2F1-pRS413) or fragment 3'-DHM2 (oligo M2F4-pRS413) were synthesized at Integrated DNA Technologies, Inc. The final Intervening DHFR module 02, plasmid pRS413-DHFR-Mod-02-Kan, was created by ligating a 2.4 kb Kan/NeoR cassette derived from DraI digestion of plasmid pEGFP-C1, with the plasmid pRS413-DHFR-Module-02 linearized by DraIII and blunted by DNA Polymerase I, Large (Klenow) Fragment.

For BAC recombineering an 11.7 kb of targeting construct was amplified from plasmid pRS413-DHFR-Mod-02-Kan using primer pair M2F12-AgeIFor/DH2-4rev and purified by gel extraction after DpnI digestion of the template plasmid.

Construction of plasmid pRS413-DHFR-Mod-03-Kan (Intervening DHFR module 03): Plasmid pRS413-DHFR-Mod-03 was made by assembling the vector backbone with four additional fragments using the yeast DNA assembler method. Fragment 5'-DHM3 (6.5 kb) and fragment 3'-DHM3 (5.0 kb) with an overlap of 1553 bp were both PCR amplified from the DHFR BAC using primer pair M3F1-AgeIFor/M3F3-BamHIrev or M3-F3For/M3F4-AgeIRev, respectively. Two bridging oligomers, with a 125 bp homology to the pRS413 vector backbone, and a 125 bp homology to fragment

5’-DHM3 (oligo M3F1-pRS413) or to fragment 3’-DHM3 (oligo M3F4-pRS413), respectively, were synthesized at Integrated DNA Technologies, Inc. The final Intervening DHFR module 03, plasmid pRS413-DHFR-Mod-03-Kan, was created by ligating a 2.4 kb Kan/NeoR cassette derived from DraI digestion of plasmid pEGFP-C1, with the plasmid pRS413-DHFR-Mod-03 linearized by SmaI.

For BAC recombineering a 12.2 kb targeting construct was amplified from plasmid pRS413-DHFR-Mod-03-Kan using primer pair DH3-1for/DH3-4rev and purified by gel extraction after DpnI digestion of the template plasmid.

Construction of plasmid pRS413-DHFR-Mod-04-Zeo (Intervening DHFR module 04): Plasmid pRS413-DHFR-Mod-04 was made by assembling the vector backbone plus 5 additional fragments using the yeast DNA assembler method (4). Fragment 5’-DHM4 (4.9 kb), fragment Mid-DHM4 (5.2 kb) and fragment 3’-DHM4 (5.2 kb) with an overlap of 2663 bp in between 5’-DHM4 and Mid-DHM4, and an overlap of 2542 bp in between Mid-DHM4 and 3’-DHM4, were PCR amplified from the DHFR BAC using primer pair M4F1-AgeIfor/DHM4F2-R, DHM4F2-Fw/DHM4F3-R, or Fw-M4F2-BamHI/RevM4F5-MluI, respectively. Two bridging oligomers, with a 125 bp homology to pRS413 vector backbone, and a 125 bp homology to fragment 5’-DHM4 (oligo M4F1-pRS413), or to fragment 3’-DHM4 (oligo M4F5-pRS413), were synthesized at Integrated DNA Technologies, Inc. The final Intervening DHFR module 04, plasmid pRS413-DHFR-Mod-04-Zeo, was created by ligating a 1.1 kb ZeoR expression cassette PCR amplified from plasmid pSV40/Zeo2 (ThermoFisher Scientific) using 5’ phosphorylated primer pair ZeoMluIFor/ZeoMluIRev, with the plasmid pRS413-DHFR-Module-04 linearized by BmgBI.

For BAC recombineering an 11.6 kb targeting construct was excised out from plasmid pRS413-DHFR-Mod-04-Zeo using KpnI/DrdI restriction enzymes and gel purified.

Assembly of modules to create multi-reporter DHFR BAC: The six targeting constructs derived from the three reporter modules and the three intervening DHFR modules were incorporated into the DHFR BAC by BAC recombineering, with the following order: Reporter module 01, Intervening DHFR module 02, Reporter module 02, Intervening DHFR module 03, Reporter module 03 and Intervening DHFR module 04. *E. coli* strain SW102 was used for BAC recombineering. Each round of BAC recombineering used a corresponding antibiotic (50 μg/ml Kanamycin, 50 μg/ml Spectinomycin, or 25 μg/ml Zeocin) as positive selection for incorporation of the current targeting construct as described in section "Construction of dual reporter DHFR BACs". In the second to the last round of BAC recombineering, colonies were further screened for loss of the antibiotic resistance gene incorporated in the previous round of BAC recombineering by streaking colonies onto a plate containing the corresponding antibiotic. Each round of recombination was validated by restriction enzyme fingerprinting.

Data availability

The raw reads of the WGS data were deposited at the Sequence Read Archive (SRA) database (BioProject number: PRJNA553146) at the National Center for Biotechnology Information (NCBI).

Code availability

All computational scripts used for CNV analysis are available at https://bitbucket.org/Binhui/bz_cnv_analysis/src.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences [GM098319 to A.S.B. and in part GM58460 to A.S.B]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

We thank Edith Heard (Curie Institute) for providing DHFR BAC (clone 057L22 from CITB mouse library), Veena K Parnaik (CSIR-CCMB, Hyderabad, India) for GFP-Lamin B1 plasmid, Miroslav Dundr (Rosalind Franklin University of Medicine and Science) for GFP-Fibrillarin plasmid, Huimin Zhao (University of Illinois Urbana-Champaign) for pQCXIN-TetR-mCherry plasmid, Peter Adams (Sanford Burnham Prebys Medical Discovery Institute) for BJ-hTERT cells, and KV Prasanth (University of Illinois Urbana-Champaign) for mRFP-Magoh plasmid. Use of the BD FACS AriaII was assisted by the flow cytometry facility stuff at Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign (UIUC).

FIGURES AND TABLES

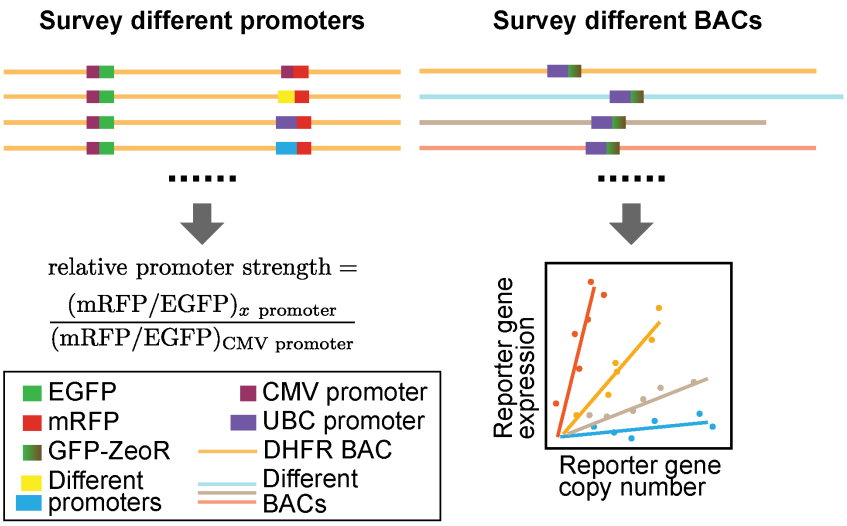

Figure 2.1. Two-prong experimental approach. Left: Identification of promoters of different strengths- We measured relative promoter strengths by embedding EGFP and mRFP reporter genes into the DHFR BAC, using the CMV promoter to drive EGFP expression and the test promoter to drive mRFP. The ratio of mRFP and GFP expression, normalized by this same ratio for a CMV test promoter, defines promoter strength relative to CMV. Right: Surveying reporter gene expression in different BAC scaffolds- (Top) The UBC-GFP-ZeoR reporter gene was inserted into BACs carrying DNA from mouse or human genomic regions corresponding to either transcriptionally active or inactive genomic regions. (Bottom) Plotting reporter gene expression (y-axis) versus reporter gene copy number (x-axis) for multiple cell clones stably expressing BAC transgenes: a linear correlation would indicate copy-number dependent, position independent expression, while the slope of this linear correlation would measure reporter gene expression per copy number.

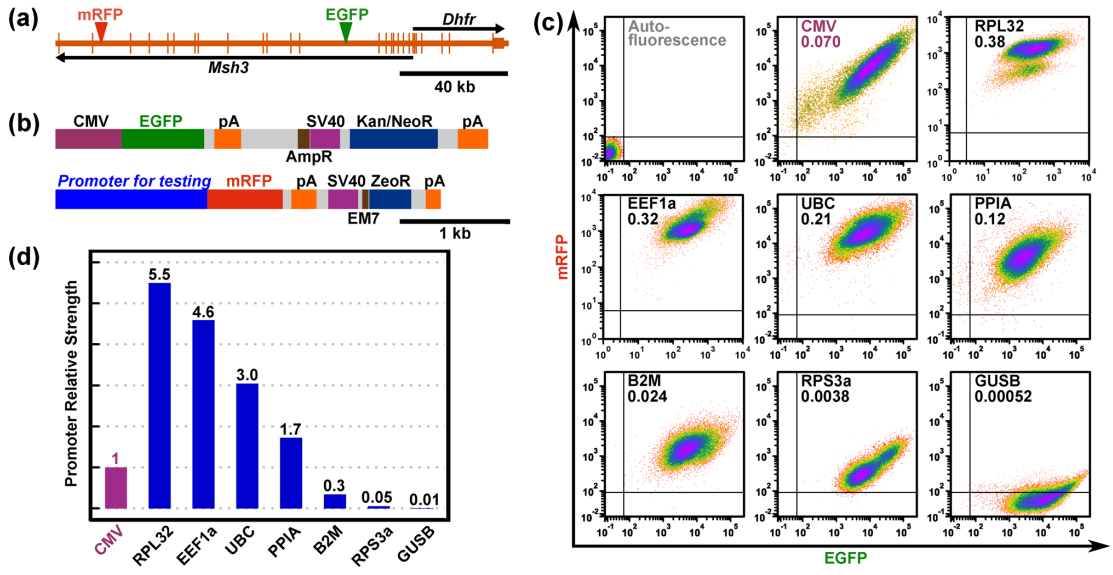

Figure 2.2 Dual-reporter assay for promoter strength estimation. (a) Dual reporter DHFR BAC showing the two genes on the BAC, *Dhfr* and *Msh3*, and the insertion sites of the two reporter expression cassettes. Longer vertical bars- exons; shorter vertical bars- UTRs; arrows- direction of transcription; green arrowhead- EGFP expression cassette insertion site; red arrowhead- mRFP expression cassette insertion site. (b) The two reporter gene/selectable marker cassettes used in the assay. The EGFP cassette (top) contains an EGFP minigene, driven by a CMV promoter, and a Kanamycin/Neomycin resistance gene (Kan/NeoR), driven by a SV40 promoter for expression in mammalian cells, or by a AmpR promoter for expression in bacteria. The mRFP cassette (bottom) contains a mRFP minigene and a Zeocin resistance gene (ZeoR). Different endogenous promoters were inserted immediately upstream of mRFP. ZeoR is driven by a SV40 promoter for expression in mammalian cells, or by a AmpR promoter for expression in bacteria. pA- poly(A) signal. (c) Scatter plots showing mRFP fluorescence (y-axis) vs EGFP fluorescence (x-axis) of cells from the mixed clonal populations stably transfected with dual reporter DHFR BACs. Promoters driving the mRFP and the ratio of mRFP/EGFP (promoter strength) are labeled in each plot. (d) Promoter strengths relative to CMV.

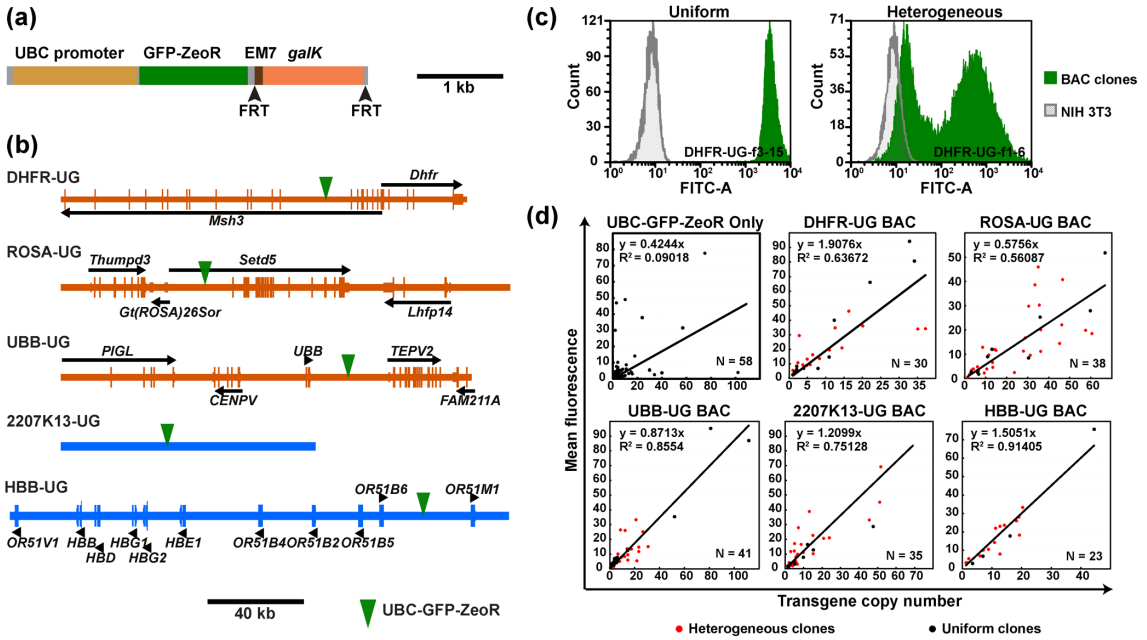

Figure 2.3. Expression of reporter gene embedded in different BAC scaffolds. (a) UBC-GFP-ZeoR-FRT-GalK-FRT cassette showing the GFP-ZeoR minigene driven by the UBC promoter and the *galK* positive/negative selection marker flanked by 34 bp flippase recognition target (FRT) sites (arrowheads). (b) Maps of the BACs used in the study. Longer vertical bars- exons; shorter vertical bars- UTRs; black arrows or arrowheads- direction of transcription; green arrow heads- UBC-GFP-ZeoR insertion site. (c) GFP fluorescence histograms obtained by flow-cytometry for "uniform" (left, green, clone DHFR-UG-f3-15) versus "heterogeneous" (right, green, clone DHFR-UG-f1-6) expressing NIH 3T3 clones carrying the DHFR-UG BAC. x-axis- fluorescence value, y-axis- cell number; gray- autofluorescence of untransfected cells. Fluorescence is measured in arbitrary units. (d) Scatter plots of mean normalized cellular GFP fluorescence (y-axis) vs reporter gene copy number (x-axis) for clonal populations transfected with the UBC-GFP-ZeoR cassette alone or with different BAC scaffolds carrying the UBC-GFP-ZeoR reporter gene. Linear regression fits (black lines, y-intercepts set to 0) are shown with corresponding R-squared values and equations. Red circles- heterogeneous clones; Black circles- uniform clones; Bottom right of plots: Number of clones analyzed.

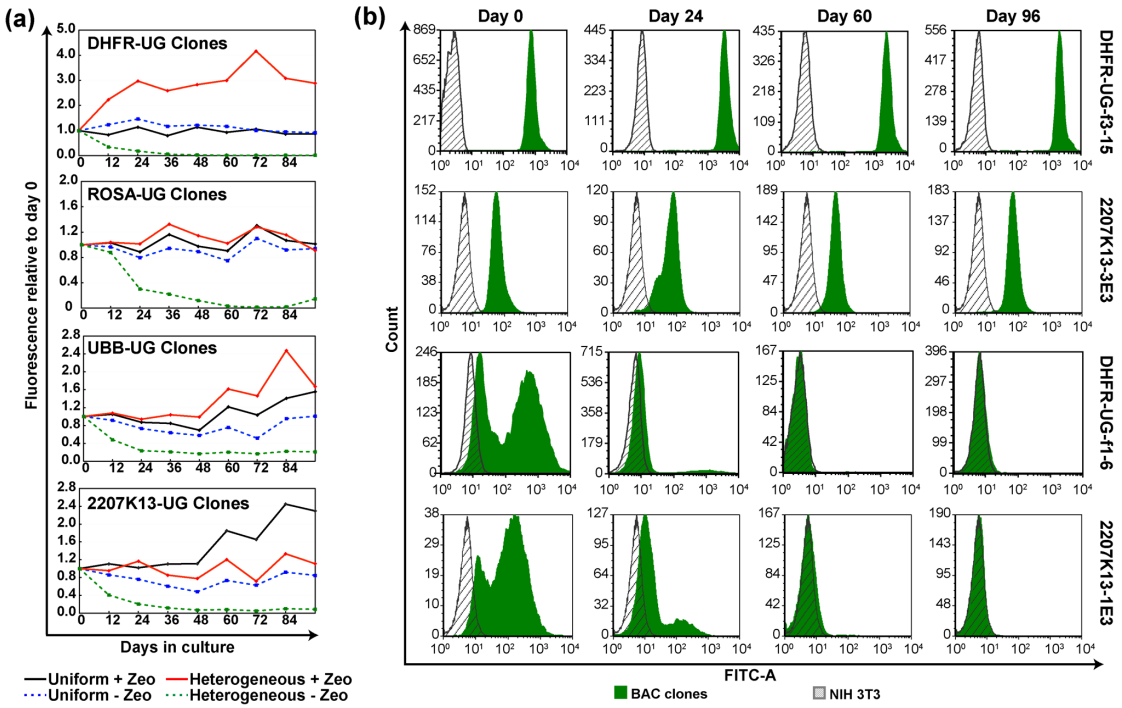

Figure 2.4. UBC-GFP-ZeoR reporter gene expression over time. "Uniform" clones show stable expression with or without expression, while "heterogenous" clones show progressive loss of expression without selection. (a) Changes in GFP fluorescence of uniform versus heterogeneous clones, averaged over multiple clones (2-8), carrying indicated BAC transgenes during 96 days of continuous passaging with or without Zeocin selection. x-axis- number of days since removal of Zeocin; y-axis- mean fluorescence values of multiple clones divided by that at day zero; black- "uniform" expressing clones cultured with Zeocin; blue- "uniform" expressing clones cultured without Zeocin; red- "heterogeneous" expressing clones cultured with Zeocin; green- "heterogeneous" expressing clones cultured without Zeocin; (b) GFP fluorescence histogram of representative "uniform" and "heterogeneous" expressing NIH 3T3 clones at day 0, 24, 60 and 96 without selection. Gray- autofluorescence of untransfected cells; Green- GFP fluorescence of the indicated clones. x-axis- fluorescence; y-axis- cell number.

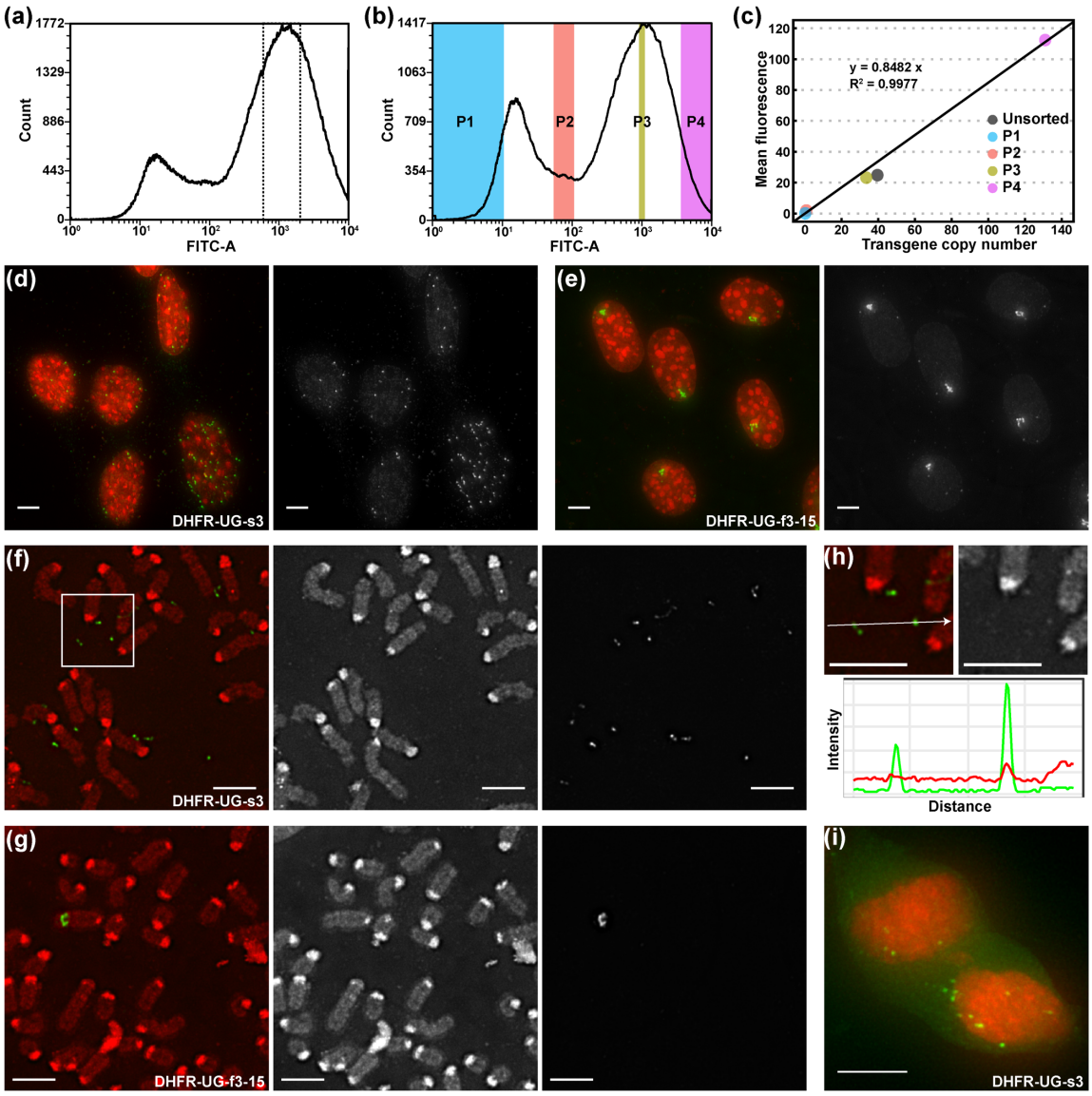

Figure 2.5. BAC transgenes exist as episomes in heterogeneously expressing clones. (a-c) BAC copy number analysis of sub-populations of a heterogeneous clone, DHFR-UG-s3, with different fluorescence levels. (a) GFP fluorescence histogram of DHFR-UG-s3 cells during first sorting (y-axis- cell number; x-axis- GFP fluorescence level). Cells within a narrow peak-window (dotted lines) were sorted by FACS. (b) GFP fluorescence histogram of sorted DHFR-UG-s3 cells after one week of cell growth. Cells within the four colored windows (P1-4) were sorted by FACS and used for BAC copy number estimation by qPCR. (c) Mean GFP fluorescence (y-axis) vs copy number (x-axis) of the four cell sub-populations and the original unsorted population shows linear correlation

Figure 2.5. Cont.

between fluorescence levels and copy number (R^2=0.99). (d-e) DNA FISH over interphase nuclei of the heterogeneous clone DHFR-UG-s3 (d) and a uniform clone DHFR-UG-f3-15 (e) to visualize the BAC transgenes. Maximum-intensity projections are shown. Gamma=0.5 was applied to FISH channel after projection to better display low intensity FISH spots. (f-g) DNA FISH over mitotic spreads of the heterogeneous clone DHFR-UG-s3 (f) and the uniform clone DHFR-UG-f3-15 (g). (h) DAPI intensity over an episome with strong FISH signal and one with weak FISH signal. Top: enlarged view of the white square area in (f); bottom: DAPI (red) and FISH signal (green) intensity profile along the white arrow in the top panel. (i) A pair of telophase nuclei of the heterogeneous clone, DHFR-UG-s3, showing unequal segregation of episomal BAC transgenes during mitosis. (d-i) Red- DNA DAPI stain; green- BAC FISH signal. Scale bars = 5 μm.

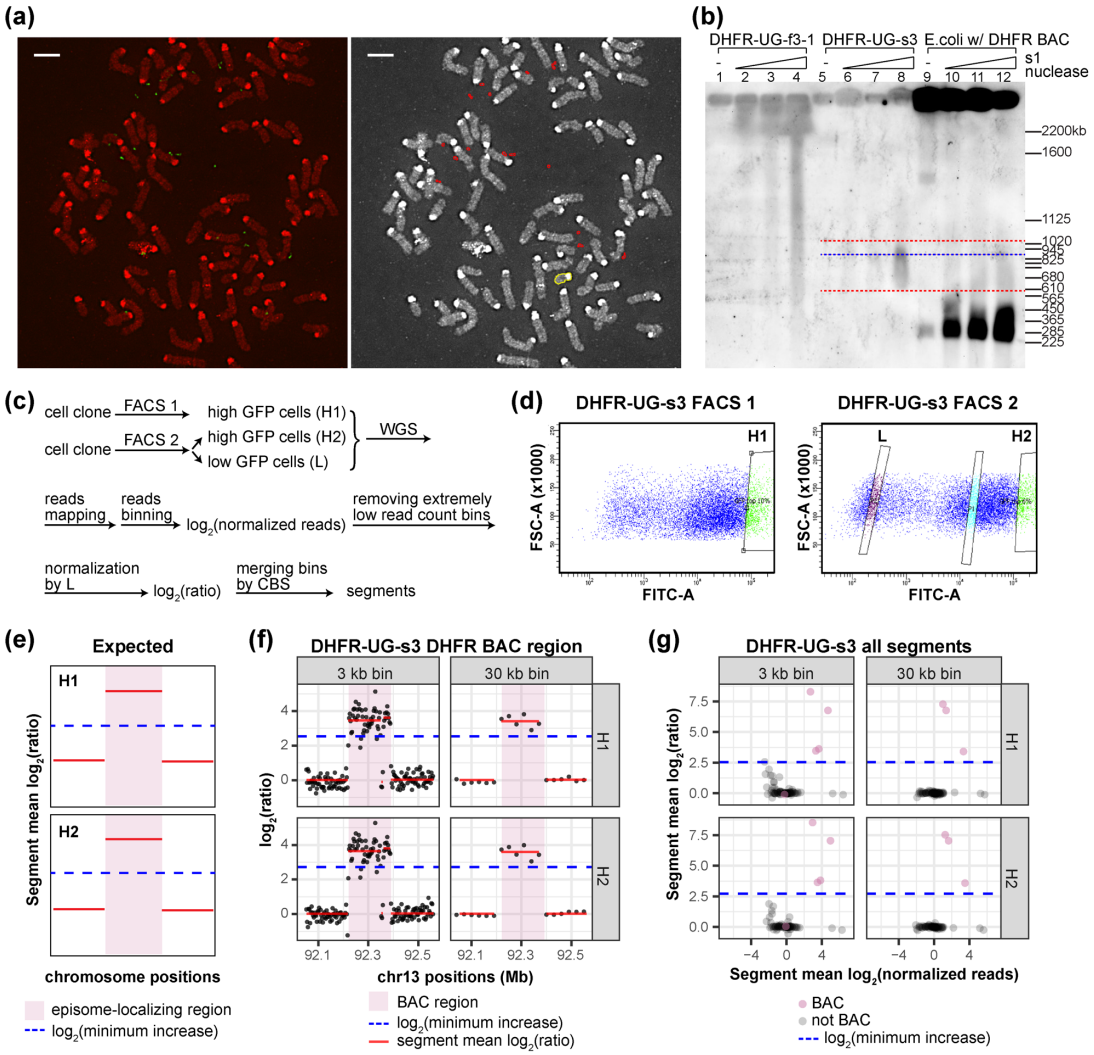

Figure 2.6. BAC episome size estimation and CNV analysis. (a) Estimation of average episome size in the DHFR-UG-s3 clone using mitotic FISH. Red- DNA DAPI stain; Green- BAC FISH signal; Red circles: regions of interest (ROIs) of FISH spots used for analysis; Yellow circles: ROI of the smallest chromosome in the field. Scale bars = 5 μm. This panel reuses the image in Figure 5f for analysis. (b) Southern hybridization using probes prepared from the DHFR BAC of cellular DNA without enzyme digestion, or digested with increasing amount of S1 Nuclease, separated by PFGE. Lane 1-4: uniform clone DHFR-UG-f3-1; Lane 5-8: heterogeneous clone DHFR-UG-s3; Lane 9-12: *E. coli* carrying the DHFR BAC. (c-g) CNV analysis of the DHFR-UG-s3 clone. (c) Flow chart of the CNV analysis. (d) Two FACS experiments for collecting cells with high (H1 and H2), and low (L) fluorescence subpopulation. x-axis- FITC channel intensity; y-axis-

Figure 2.6. Cont.

forward scatter; H1, H2, and L- sorting windows. (e) Episome-localizing genomic regions (pink highlighted regions) are expected to have mean $\log_2$(ratio) (red line) equal to or greater than $\log_2$(estimated minimum copy number increase) (blue dashed line). (f) $\log_2$(ratio) of individual bins (dark gray dots) and the segment mean $\log_2$(ratio) (red lines) around the *Dhfr-Msh3* locus belonging to the DHFR BAC (pink highlight) in the H1 and H2 subpopulations of the DHFR-UG-s3 clone. (g) Scatter plot of segment mean $\log_2$(ratio) vs segment mean $\log_2$(normalized reads) of all segments of the H1 and H2 subpopulations of the DHFR-UG-s3 clone. Pink dots- segments belonging to the DHFR BAC, including the *Dhfr-Msh3* locus, UBC-GFP-ZeoR and the BAC vector; Black dots- remaining segments in the genome. (f-g) Blue dashed line: $\log_2$(estimated minimum copy number increase).

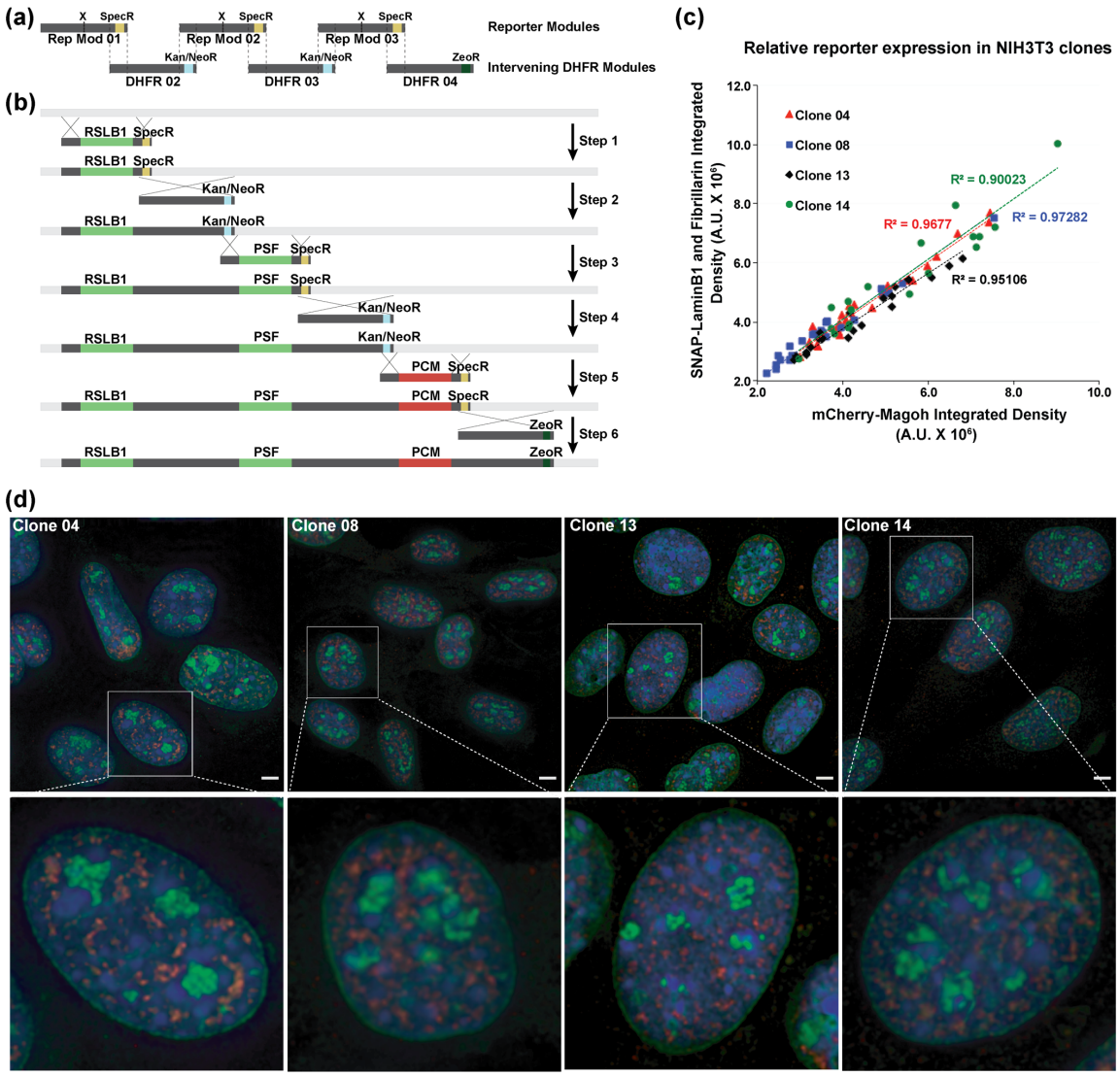

Figure 2.7. BAC-MAGIC and simultaneous multi-reporter expression. (a-b) Construction of the multi-reporter DHFR BAC by BAC-MAGIC. (a) Modular design of BAC-MAGIC. Reporter module 01, 02 and 03 contain reporter gene expression cassettes (X), DHFR BAC homologous sequences (dark gray), and Spectinomycin resistance markers (SpecR, yellow) near the 3' ends for bacterial selection. Intervening DHFR module 02, 03 and 04 contain DHFR BAC homologous sequences (dark gray), and antibiotic resistance markers near the 3' ends (Kanamycin/Neomycin resistance marker (Kan/NeoR, blue) in module 02 and 03 for bacterial selection, and Zeocin resistance marker (ZeoR, dark green) in module 04 for dual selection in bacterial or mammalian

Figure 2.7. Cont.

cells). The dotted lines mark homologous regions between the reporter modules and the intervening DHFR modules. (b) Six sequential steps of BAC recombineering introduce three reporter expression cassettes, RPL32-driven SNAP-tagged Lamin B1 (RSLB1), PPIA-driven SNAP-tagged Fibrillarin (PSF), and PPIA-driven mCherry-Magoh, onto the DHFR BAC (light gray) with ~10 kb of intervening DHFR BAC sequences (dark gray). Homologous regions are indicated by crossed lines. (c) Relative expression of the SNAP-tagged Lamin B1 and Fibrillarin to the mCherry-Magoh reporter in four representative NIH 3T3 cell clones (04, 08, 13 and 14) containing the multi-reporter BAC. Integrated fluorescence intensities per cell of SNAP--fluorescein (y-axis) and mCherry-Magoh (x-axis) are plotted. Linear regression lines (y-intercepts set to 0) are shown with corresponding R-squared values. Number of nuclei of each clone analyzed range from 18 to 27. Red- Clone 04; Blue- Clone 08, Black- Clone 13; Green- Clone 14. (d) Representative images (maximum intensity projections of 2-3 optical sections) from the four cell clones (Clone 04, 08, 13 and 14) showing expression of the three reporter genes. Nuclear lamina is labeled with SNAP-tagged Lamin B1 (green), nucleoli with SNAP-tagged Fibrillarin (green), and speckles with mCherry-Magoh (red). One magnified nucleus from each representative field (top panel) is shown in the bottom panel. Scale bars = 5 μm.

Table 2.1. Percentage of heterogeneously expressing clones transfected with the UBC-GFP-ZeoR cassette alone or with different BAC scaffolds carrying the UBC-GFP-ZeoR reporter gene.

Construct	Heterogeneous clones%	Number of clones
UBC-GFP-ZeoR	0	58
DHFR-UG	60%	30
ROSA-UG	76%	38
UBB-UG	58%	41
2207K13-UG	69%	35
HBB-UG	83%	23

Table 2.2. BAC copy number, episome copy number, and BAC DNA content per episome in clone DHFR-UG-s3 and clone HBB-UG-100d3.

Sample name	BAC copy number per cell	Episome copy number per cell	BAC copy number per episome	BAC size (kb)	BAC content per episome (kb)
DHFR-UG-s3	15.4	6.2 (n=99)	2.5	178	445
HBB-UG-100d3	14.9	4.5 (n=100)	3.3	217	716

Table 2.3. BAC copy number, estimated episome copy number, and estimated minimum copy number increase of episome-localizing DNA in H1 and H2 subpopulations relative to L subpopulation of clone DHFR-UG-s3 and clone HBB-UG-100d3.

Sample name	BAC copy number per cell	Estimated episome copy number per cell	Minimum copy number increase of episome-localizing DNA relative to L
DHFR-UG-s3_H1	49.4	19.8	5.8
DHFR-UG-s3_H2	57.5	23.1	6.6
DHFR-UG-s3_L	0.3	0.1	/
HBB-UG-100d3_H1	48.5	14.7	4.6
HBB-UG-100d3_H2	51.2	15.5	4.8
HBB-UG-100d3_L	0.2	0.1	/

REFERENCES

Akhtar, W. *et al.* (2013) 'Chromatin position effects assayed by thousands of reporters integrated in parallel.', *Cell*. Elsevier, 154(4), pp. 914–27.

Argyros, O. *et al.* (2008) 'Persistent episomal transgene expression in liver following delivery of a scaffold/matrix attachment region containing non-viral vector.', *Gene therapy*, 15(24), pp. 1593–1605.

Baiker, A. *et al.* (2000) 'Mitotic stability of an episomal vector containing a human scaffold/matrix-attached region is provided by association with nuclear matrix.', *Nature cell biology*, 2(3), pp. 182–4.

Barton, B. M., Harding, G. P. and Zuccarelli, A. J. (1995) 'A general method for detecting and sizing large plasmids.', *Analytical biochemistry*, 226(2), pp. 235–40.

Beatty, B. G. and Scherer, S. W. (2002) 'Human chromosome mapping of single copy genes', *FISH: A practical approach. B. Beatty, S. Mai, and J. Squire, editors. Oxford University Press, Oxford*, pp. 29–53.

Bertulat, B. *et al.* (2012) 'MeCP2 Dependent Heterochromatin Reorganization during Neural Differentiation of a Novel Mecp2-Deficient Embryonic Stem Cell Reporter Line', *PLoS ONE*, 7.

Bharadwaj, R. R. *et al.* (2003) 'LCR-regulated transgene expression levels depend on the Oct-1 site in the AT-rich region of beta -globin intron-2.', *Blood*, 101(4), pp. 1603–10.

Bian, Q. *et al.* (2013) 'β-Globin cis-elements determine differential nuclear targeting through epigenetic modifications.', *The Journal of cell biology*, 203(5), pp. 767–83.

Bian, Q. and Belmont, A. S. (2010) 'BAC TG-EMBED: one-step method for high-level, copy-number-dependent, position-independent transgene expression.', *Nucleic acids research*, 38(11), p. e127.

Blaas, L. *et al.* (2009) 'Bacterial artificial chromosomes improve recombinant protein production in mammalian cells', *BMC Biotechnology*, 9(1), p. 3.

Brooks, A. R. *et al.* (2004) 'Transcriptional silencing is associated with extensive methylation of the CMV promoter following adenoviral gene delivery to muscle.', *The journal of gene medicine*, 6(4), pp. 395–404.

Brophy, J. A. N. and Voigt, C. A. (2014) 'Principles of genetic circuit design', *Nature Methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 11, p. 508.

Carroll, S. M. *et al.* (1988) 'Double minute chromosomes can be produced from precursors derived from a chromosomal deletion.', *Molecular and cellular biology*, 8(4), pp. 1525–33.

Chaturvedi, P. *et al.* (2018) 'Stable and reproducible transgene expression independent of proliferative or differentiated state using BAC TG-EMBED.', *Gene therapy*, 25(5), pp. 376–391.

Chen, C. *et al.* (2011) 'A comparison of exogenous promoter activity at the ROSA26 locus using a PhiC31 integrase mediated cassette exchange approach in mouse ES cells.', *PloS one*, 6(8), p. e23376.

Chen, M. *et al.* (2013) 'Decoupling Epigenetic and Genetic Effects through Systematic Analysis of Gene Position.', *Cell reports*. Elsevier, 3(1), pp. 128–37.

Chen, Z. Y. *et al.* (2004) 'Silencing of episomal transgene expression by plasmid bacterial DNA elements in vivo.', *Gene therapy*, 11(10), pp. 856–864.

Conese, M., Auriche, C. and Ascenzioni, F. (2004) 'Gene therapy progress and prospects: episomally maintained self-replicating systems.', *Gene therapy*. Nature Publishing Group, 11(24), pp. 1735–41.

Van Craenenbroeck, K., Vanhoenacker, P. and Haegeman, G. (2000) 'Episomal vectors for gene expression in mammalian cells.', *European journal of biochemistry*, 267(18), pp. 5665–78.

Cremer, M. *et al.* (2008) 'Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes.', *Methods in molecular biology (Clifton, N.J.)*, 463, pp. 205–39.

Dernburg, A. F. (2011) 'Fragmentation and labeling of probe DNA for whole-mount FISH in Drosophila.', *Cold Spring Harbor protocols*, 2011(12), pp. 1527–30.

Dorer, D. R. and Henikoff, S. (1997) 'Transgene repeat arrays interact with distant heterochromatin and cause silencing in cis and trans.', *Genetics*, 147(3), pp. 1181–90.

Ehrhardt, A. *et al.* (2008) 'Episomal vectors for gene therapy.', *Current gene therapy*, 8(3), pp. 147–61.

Emery, D. W. *et al.* (2000) 'A chromatin insulator protects retrovirus vectors from chromosomal position effects', *Proceedings of the National Academy of Sciences*, 97(16), pp. 9150–9155.

Fitzsimons, H. L., Bland, R. J. and During, M. J. (2002) 'Promoters and regulatory elements that improve adeno-associated virus transgene expression in the brain.', *Methods (San Diego, Calif.)*, 28(2), pp. 227–36.

Fournier, R. E. and Ruddle, F. H. (1977) 'Microcell-mediated transfer of murine chromosomes into mouse, Chinese hamster, and human somatic cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 74(1), pp. 319–23.

Frappier, L. (2012) ‘Contributions of Epstein-Barr nuclear antigen 1 (EBNA1) to cell immortalization and survival.’, *Viruses*, 4(9), pp. 1537–47.

Garrick, D. *et al.* (1998) ‘Repeat-induced gene silencing in mammals.’, *Nature genetics*, 18(1), pp. 56–9.

Gibson, D. G. *et al.* (2009) ‘Enzymatic assembly of DNA molecules up to several hundred kilobases.’, *Nature methods*, 6(5), pp. 343–5.

Gietz, R. D. and Schiestl, R. H. (2007) ‘Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method.’, *Nature protocols*, 2(1), pp. 38–41.

Girton, J. R. and Johansen, K. M. (2008) ‘Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila.’, *Advances in genetics*, 61(07), pp. 1–43.

Glover, D. J., Lipps, H. J. and Jans, D. A. (2005) ‘Towards safe, non-viral therapeutic gene expression in humans.’, *Nature reviews. Genetics*, 6(4), pp. 299–310.

Grandchamp, N. *et al.* (2011) ‘Influence of insulators on transgene expression from integrating and non-integrating lentiviral vectors.’, *Genetic vaccines and therapy*, 9(1), p. 1.

Grosveld, F. *et al.* (1987) ‘Position-independent, high-level expression of the human beta-globin gene in transgenic mice.’, *Cell*, 51(6), pp. 975–85.

Guy, L. G. *et al.* (1996) ‘The beta-globin locus control region enhances transcription of but does not confer position-independent expression onto the lacZ gene in transgenic mice.’, *The EMBO journal*, 15(14), pp. 3713–21.

Harrington, J. J. *et al.* (1997) ‘Formation of de novo centromeres and construction of first-generation human artificial microchromosomes.’, *Nature genetics*, 15(4), pp. 345–55.

He, J., Yang, Q. and Chang, L.-J. (2005) ‘Dynamic DNA Methylation and Histone Modifications Contribute to Lentiviral Transgene Silencing in Murine Embryonic Carcinoma Cells’, *Journal of Virology*, 79(21), pp. 13497–13508.

Herbst, F. *et al.* (2012) ‘Extensive methylation of promoter sequences silences lentiviral transgene expression during stem cell differentiation in vivo.’, *Molecular therapy : the journal of the American Society of Gene Therapy*, 20(5), pp. 1014–21.

Hiratsuka, M. *et al.* (2011) ‘Integration-free iPS cells engineered using human artificial chromosome vectors.’, *PloS one*, 6(10), p. e25961.

Hong Cai, J. *et al.* (2007) ‘Validation of rat reference genes for improved quantitative gene expression analysis using low density arrays’, *BioTechniques*, 42(4), pp. 503–512.

Hong, S. *et al.* (2007) 'Functional analysis of various promoters in lentiviral vectors at different stages of in vitro differentiation of mouse embryonic stem cells.', *Molecular therapy : the journal of the American Society of Gene Therapy*, 15(9), pp. 1630–9.

Hotta, A. and Ellis, J. (2008) 'Retroviral vector silencing during iPS cell induction: an epigenetic beacon that signals distinct pluripotent states.', *Journal of cellular biochemistry*, 105(4), pp. 940–8.

Hu, Y. *et al.* (2009) 'Large-scale chromatin structure of inducible genes: transcription on a condensed, linear template', *The Journal of Cell Biology*, 185(1), pp. 87–100.

Hu, Y., Plutz, M. and Belmont, A. S. (2010) 'Hsp70 gene association with nuclear speckles is Hsp70 promoter specific.', *The Journal of cell biology*, 191(4), pp. 711–9.

Jenke, A. C. W. *et al.* (2004) 'Nuclear scaffold/matrix attached region modules linked to a transcription unit are sufficient for replication and maintenance of a mammalian episome', *Proceedings of the National Academy of Sciences*, 101(31), pp. 11322–11327.

de Jonge, H. J. M. *et al.* (2007) 'Evidence based selection of housekeeping genes.', *PloS one*, 2(9), p. e898.

Karpen, G. H. (1994) 'Position-effect variegation and the new biology of heterochromatin.', *Current opinion in genetics & development*, 4(2), pp. 281–91.

Kazuki, Y. *et al.* (2010) 'Complete genetic correction of ips cells from Duchenne muscular dystrophy.', *Molecular therapy : the journal of the American Society of Gene Therapy*, 18(2), pp. 386–93.

Kazuki, Y. and Oshimura, M. (2011) 'Human artificial chromosomes for gene delivery and the development of animal models.', *Molecular therapy : the journal of the American Society of Gene Therapy*, 19(9), pp. 1591–601.

Khan, S. R. and Kuzminov, A. (2017) 'Degradation of RNA during lysis of Escherichia coli cells in agarose plugs breaks the chromosome.', *PloS one*, 12(12), p. e0190177.

Khanna, N. *et al.* (2013) 'BAC manipulations for making BAC transgene arrays.', *Methods in molecular biology (Clifton, N.J.)*, 1042, pp. 197–210.

Khanna, N., Hu, Y. and Belmont, A. S. S. (2014) 'HSP70 Transgene Directed Motion to Nuclear Speckles Facilitates Heat Shock Activation', *Current Biology*. Elsevier Ltd, 24(10), pp. 1138–1144.

Kim, J.-H. *et al.* (2011) 'Human artificial chromosome (HAC) vector with a conditional centromere for correction of genetic deficiencies in human cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp. 20048–53.

Kim, J. M. *et al.* (2004) 'Improved recombinant gene expression in CHO cells using matrix attachment regions', *Journal of biotechnology*, 107(2), pp. 95–105.

Kimura, M. *et al.* (2010) ‘Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths.’, *Nature protocols*, 5(9), pp. 1596–607.

Kouprina, N. *et al.* (2013) ‘A new generation of human artificial chromosomes for functional genomics and gene therapy.’, *Cellular and molecular life sciences : CMLS*, 70(7), pp. 1135–48.

Kwaks, T. H. J. *et al.* (2003) ‘Identification of anti-repressor elements that confer high and stable protein production in mammalian cells.’, *Nature biotechnology*, 21(5), pp. 553–8.

L’Abbate, A. *et al.* (2014) ‘Genomic organization and evolution of double minutes/homogeneously staining regions with MYC amplification in human cancer.’, *Nucleic acids research*, 42(14), pp. 9131–45.

Laker, C. *et al.* (1998) ‘Host cis-mediated extinction of a retrovirus permissive for expression in embryonal stem cells during differentiation.’, *Journal of virology*, 72(1), pp. 339–348.

Larin, Z. and Mejía, J. E. (2002) ‘Advances in human artificial chromosome technology’, *Trends in Genetics*, 18(6), pp. 313–319.

Liskovykh, M. *et al.* (2016) ‘Moving toward a higher efficiency of microcell-mediated chromosome transfer.’, *Molecular therapy. Methods & clinical development*, 3, p. 16043.

Lois, C. *et al.* (2002) ‘Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors.’, *Science (New York, N.Y.)*, 295(5556), pp. 868–72.

Lufino, M. M. P., Edser, P. A. H. and Wade-Martins, R. (2008) ‘Advances in high-capacity extrachromosomal vector technology: Episomal maintenance, vector delivery, and transgene expression’, *Molecular Therapy*. The American Society of Gene Therapy, 16(9), pp. 1525–1538.

Machida, K. *et al.* (2012) ‘Reconstitution of the human chaperonin CCT by co-expression of the eight distinct subunits in mammalian cells’, *PROTEIN EXPRESSION AND PURIFICATION*. Elsevier Inc., 82(1), pp. 61–69.

Marasini, D. and Fakhr, M. (2014) ‘Exploring PFGE for Detecting Large Plasmids in Campylobacter jejuni and Campylobacter coli Isolated from Various Retail Meats’, *Pathogens*, 3(4), pp. 833–844.

Mills, W. *et al.* (1999) ‘Generation of an approximately 2.4 Mb human X centromere-based minichromosome by targeted telomere-associated chromosome fragmentation in DT40.’, *Human molecular genetics*, 8(5), pp. 751–61.

Minoguchi, S. and Iba, H. (2008) ‘Instability of retroviral DNA methylation in embryonic stem cells.’, *Stem cells (Dayton, Ohio)*, 26(5), pp. 1166–73.

Mizushima, S. and Nagata, S. (1990) ‘pEF-BOS, a powerful mammalian expression vector.’, *Nucleic acids research*, 18(17), p. 5322.

‘Mowiol mounting medium’ (2006) *Cold Spring Harbor Protocols*, 2006(1), p. pdb.rec10255.

Müller-Kuller, U. *et al.* (2015) ‘A minimal ubiquitous chromatin opening element (UCOE) effectively prevents silencing of juxtaposed heterologous promoters by epigenetic remodeling in multipotent and pluripotent stem cells.’, *Nucleic acids research*, 43(3), pp. 1577–92.

Nanbo, A., Sugden, A. and Sugden, B. (2007) ‘The coupling of synthesis and partitioning of EBV’s plasmid replicon is revealed in live cells.’, *The EMBO journal*, 26(19), pp. 4252–62.

Olshen, A. B. *et al.* (2004) ‘Circular binary segmentation for the analysis of array-based DNA copy number data’, *Biostatistics*, 5(4), pp. 557–572.

Phi-Van, L. *et al.* (1990) ‘The chicken lysozyme 5’ matrix attachment region increases transcription from a heterologous promoter in heterologous cells and dampens position effects on the expression of transfected genes.’, *Molecular and cellular biology*, 10(5), pp. 2302–7.

Piechaczek, C. *et al.* (1999) ‘A vector based on the SV40 origin of replication and chromosomal S/MARs replicates episomally in CHO cells’, *Nucleic Acids Research*, 27(2), pp. 426–428.

Pikaart, M. J., Recillas-Targa, F. and Felsenfeld, G. (1998) ‘Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators’, *Genes & Development*, 12(18), pp. 2852–2862.

Prelich, G. (2012) ‘Gene overexpression: uses, mechanisms, and interpretation.’, *Genetics*, 190(3), pp. 841–54.

Qin, J. Y. *et al.* (2010) ‘Systematic comparison of constitutive promoters and the doxycycline-inducible promoter.’, *PloS one*, 5(5), p. e10611.

Ramunas, J. *et al.* (2007) ‘Real-time fluorescence tracking of dynamic transgene variegation in stem cells.’, *Molecular therapy : the journal of the American Society of Gene Therapy*, 15(4), pp. 810–7.

Richardson, S. M. *et al.* (2017) ‘Design of a synthetic yeast genome.’, *Science (New York, N.Y.)*, 355(6329), pp. 1040–1044.

Riu, E. *et al.* (2007) ‘Histone modifications are associated with the persistence or silencing of vector-mediated transgene expression in vivo.’, *Molecular therapy : the journal of the American Society of Gene Therapy*, 15(7), pp. 1348–55.

Rizzo, M. A., Davidson, M. W. and Piston, D. W. (2009) 'Fluorescent Protein Tracking and Detection: Applications Using Fluorescent Proteins in Living Cells', *Cold Spring Harbor Protocols*, 2009(12), p. pdb.top64-pdb.top64.

Robertson, G. *et al.* (1995) 'Position-dependent variegation of globin transgene expression in mice.', *Proceedings of the National Academy of Sciences of the United States of America*, 92(12), pp. 5371–5.

Rozen, S. and Skaletsky, H. (2000) 'Primer3 on the WWW for general users and for biologist programmers.', *Methods in molecular biology (Clifton, N.J.)*, 132, pp. 365–86.

Sambrook, J. and Russell, D. W. (2006a) 'Preparation of DNA for Pulsed-field Gel Electrophoresis: Isolation of DNA from Mammalian Cells and Tissues', *Cold Spring Harbor Protocols*. Edited by J. Sambrook and D. Russell, 2006(1), p. pdb.prot4030.

Sambrook, J. and Russell, D. W. (2006b) 'Purification of Nucleic Acids by Extraction with Phenol:Chloroform', *Cold Spring Harbor Protocols*, 2006(1), p. pdb.prot4455.

Sambrook, J. and Russell, D. W. (2006c) 'Southern Hybridization of Radiolabeled Probes to Nucleic Acids Immobilized on Membranes', *Cold Spring Harbor Protocols*. Edited by J. Sambrook and D. Russell, 2006(1), p. pdb.prot4044.

Schindelin, J. *et al.* (2012) 'Fiji: an open-source platform for biological-image analysis.', *Nature methods*, 9(7), pp. 676–82.

Schoenlein, P. V *et al.* (1992) 'Double minute chromosomes carrying the human multidrug resistance 1 and 2 genes are generated from the dimerization of submicroscopic circular DNAs in colchicine-selected KB carcinoma cells', *Mol Biol Cell*, 3(5), pp. 507–520.

Schorpp, M. *et al.* (1996) 'The Human Ubiquitin C Promoter Directs High Ubiquitous Expression of Transgenes in Mice', *Nucleic Acids Research*, 24(9), pp. 1787–1788.

Schwab, M. and Amler, L. C. (1990) 'Amplification of cellular oncogenes: A predictor of clinical outcome in human cancer', *Genes, Chromosomes and Cancer*, 1(3), pp. 181–193.

Scrable, H. and Stambrook, P. J. (1999) 'A genetic program for deletion of foreign DNA from the mammalian genome.', *Mutation research*, 429(2), pp. 225–37.

Shao, Z. and Zhao, H. (2012) *DNA assembler: A synthetic biology tool for characterizing and engineering natural product gene clusters*. 1st edn, *Methods in Enzymology*. 1st edn. Elsevier Inc.

Shao, Z., Zhao, H. H. and Zhao, H. H. (2009) 'DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways.', *Nucleic acids research*, 37(2), p. e16.

She, X. *et al.* (2009) 'Definition, conservation and epigenetics of housekeeping and tissue-enriched genes.', *BMC genomics*, 10, p. 269.

Shimizu, N. *et al.* (2001) 'Plasmids with a mammalian replication origin and a matrix attachment region initiate the event similar to gene amplification', *Cancer Research*, 61(19), pp. 6987–6990.

Shimizu, N. *et al.* (2003) 'Amplification of plasmids containing a mammalian replication initiation region is mediated by controllable conflict between replication and transcription', *Cancer Research*, 63(17), pp. 5281–5290.

Shimizu, N. (2009) 'Extrachromosomal double minutes and chromosomal homogeneously staining regions as probes for chromosome research', *Cytogenetic and Genome Research*, 124(3–4), pp. 312–326.

Shimizu, N., Shingaki, K. and Kaneko-sasaguri, Y. (2005) 'When , where and how the bridge breaks : anaphase bridge breakage plays a crucial role in gene amplification and HSR generation', 302, pp. 233–243.

Sinclair, P. *et al.* (2010) 'Dynamic plasticity of large-scale chromatin structure revealed by self-assembly of engineered chromosome regions.', *The Journal of cell biology*, 190(5), pp. 761–76.

Solovei, I. and Cremer, M. (2010) '3D-FISH on cultured cells combined with immunostaining.', *Methods in molecular biology (Clifton, N.J.)*, 659, pp. 117–26.

Stehle, I. M. *et al.* (2007) 'Establishment and mitotic stability of an extra-chromosomal mammalian replicon', *BMC Cell Biology*, 8, pp. 1–12.

Strukov, Y. G. and Belmont, a. S. (2008) 'Development of Mammalian Cell Lines with lac Operator-Tagged Chromosomes', *Cold Spring Harbor Protocols*, 2008(2), p. pdb.prot4903-pdb.prot4903.

Suzuki, M., Kasai, K. and Saeki, Y. (2006) 'Plasmid DNA sequences present in conventional herpes simplex virus amplicon vectors cause rapid transgene silencing by forming inactive chromatin.', *Journal of virology*, 80(7), pp. 3293–300.

Takahashi, K. and Yamanaka, S. (2006) 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors', *Cell*, 126(4), pp. 663–676.

Takahashi, Y. *et al.* (2010) 'Development of evaluation system for bioactive substances using human artificial chromosome-mediated osteocalcin gene expression.', *Journal of biochemistry*, 148(1), pp. 29–34.

Tchasovnikarova, I. A. *et al.* (2015) 'Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells', *Science*, 348(6242), pp. 1481–1485.

Tessadori, F. *et al.* (2010) 'Stable S/MAR-based episomal vectors are regulated at the chromatin level.', *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 18(7), pp. 757–75.

Truffinet, V. *et al.* (2005) 'The chicken beta-globin HS4 insulator is not a silver bullet to obtain copy-number dependent expression of transgenes in stable B cell transfectants.', *Immunology letters*, 96(2), pp. 303–4.

Venkatraman, E. S. and Olshen, A. B. (2007) 'A faster circular binary segmentation algorithm for the analysis of array CGH data', *Bioinformatics*, 23(6), pp. 657–663.

Walker, P. R., LeBlanc, J. and Sikorska, M. (1997) 'Evidence that DNA fragmentation in apoptosis is initiated and propagated by single-strand breaks', *Cell Death and Differentiation*, 4(6), pp. 506–515.

Walsh, G. (2018) 'Biopharmaceutical benchmarks 2018.', *Nature biotechnology*, 36(12), pp. 1136–1145.

Warming, S. *et al.* (2005) 'Simple and highly efficient BAC recombineering using galK selection.', *Nucleic acids research*, 33(4), p. e36.

Williams, S. *et al.* (2005) 'CpG-island fragments from the HNRPA2B1/CBX3 genomic locus reduce silencing and enhance transgene expression from the hCMV promoter/enhancer in mammalian cells.', *BMC biotechnology*, 5, p. 17.

Wurm, F. M. (2004) 'Production of recombinant protein therapeutics in cultivated mammalian cells.', *Nature biotechnology*, 22(11), pp. 1393–8.

Ye, J. *et al.* (2012) 'Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction', *BMC Bioinformatics*, 13(1), p. 134.

Zambrowicz, B. P. *et al.* (1997) 'Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 94(8), pp. 3789–94.

Zboray, K. *et al.* (2015) 'Heterologous protein production using euchromatin-containing expression vectors in mammalian cells.', *Nucleic acids research*, 43(16), p. e102.

Zhu, J. *et al.* (2008) 'How many human genes can be defined as housekeeping with current expression data?', *BMC genomics*, 9, p. 172.

Zhu, J. (2012) 'Mammalian cell protein expression for biopharmaceutical production', *Biotechnology Advances*, 30(5), pp. 1158–1170.

CHAPTER 3: BAC TRANSGENE ARRAYS AS A MODEL SYSTEM FOR DISSECTING LARGE-SCALE CHROMATIN ORGANIZATION

ABSTRACT

Dissecting mechanisms regulating large-scale chromatin organization in mammalian cells is difficult largely due to the complexity of the mammalian genome and the lack of efficient ways both to manipulate the sequence and then to visualize the consequences of these manipulations on chromatin folding and nuclear localization. Here we show how bacterial artificial chromosomes (BACs) may be a powerful tool to dissect determinants of large-scale chromatin organization. We first show distinctive chromatin structures formed by Mb sized BAC transgene arrays containing transcriptionally active versus inactive genomic inserts. Such BAC transgene arrays have much lower DNA complexity than endogenous genomic regions of comparable size. We then show interesting clues about determinants of large-scale chromatin organization. First, insertion of a reporter mini-gene into a BAC containing the human β-globin locus (HBB BAC) moved the BAC transgene away from the nuclear periphery, without opening up the condensed chromatin formed by the HBB BAC transgene. Second, a deletion DHFR BAC, derived from a BAC containing the mouse *Dhfr-Msh3* locus (DHFR BAC) formed more condensed chromatin, but had higher transcriptional activities, compared to another deletion DHFR BAC. These results suggest separable pathways controlling large-scale chromatin compaction, nuclear localization and transcriptional activity.

INTRODUCTION

In eukaryotes, DNA is packaged with histones and other proteins, as wells as RNAs, to form chromatin in interphase cells. Large-scale chromatin organization refers to the chromatin structure beyond the "beads-on-a-string" chromatin fiber (Olins and Olins, 1974), and the spatial organization of the chromatin. While the higher levels of chromatin folding has not been solved (Bian and Belmont, 2012; Ghirlando and Felsenfeld, 2013), it has long been observed that chromatin fibers in the interphase nuclei have various diameters, as shown by transmission electron microscopy (TEM) studies (Belmont *et al.*, 1989; Bohrmann and Kellenberger, 1994; Bazett-Jones and Hendzel, 1999) and recently ChromEMT (Ou *et al.*, 2017). However, it is not clear whether such differential compaction of chromatin is sequence specific, and if so, what sequences, and how these sequences regulate chromatin compaction.

The chromatin has highly ordered spatial organization. Each chromosome occupies a distinct territory (Cremer and Cremer, 2001; Parada *et al.*, 2002). The interaction pattern of genomic regions on each chromosome form topologically associating domains (Dixon *et al.*, 2012; Dixon, Gorkin and Ren, 2016). The positioning of the genomic regions in relative to the nuclear lamina and various nuclear bodies, such as nuclear speckles, nucleoli, Cajal bodies, etc., is also non-random, as shown by both microscopy (Shopland *et al.*, 2003; Wang *et al.*, 2016) and genomic mapping studies (Pickersgill *et al.*, 2006; Guelen *et al.*, 2008; Németh *et al.*, 2010; Peric-Hupkes *et al.*, 2010; van Koningsbruggen *et al.*, 2010; Chen *et al.*, 2018; Quinodoz *et al.*, 2018). However, like with large-scale chromatin structure, the sequences and mechanisms that regulate the spatial organization of the chromatin remains to be elucidated.

Moreover, numerous studies have shown correlations between chromatin compaction, nuclear localization, epigenetic modifications, and transcriptional activities. The nuclear lamina and nucleoli are lined with condensed chromatin (Fawcett, 1966; Schöfer and Weipoltshammer, 2018). Lamina associated genomic regions (LADs) (Pickersgill *et al.*, 2006; Guelen *et al.*, 2008) are generally low in transcriptional activity, and are enriched with repressive histone modifications (Guelen *et al.*, 2008; Peric-Hupkes *et al.*, 2010). Chromatin with different combinations of epigenetic marks forms distinctive structures (Boettiger *et al.*, 2016; Xu *et al.*, 2018). Genomic regions with higher transcriptional activities are less condensed and further away from the nuclear periphery than that with lower transcriptional activities (Chambeyron and Bickmore, 2004; Gierman *et al.*, 2007; Hu *et al.*, 2009; Peric-Hupkes *et al.*, 2010; Lund *et al.*, 2013; Robson *et al.*, 2016). However, whether these correlations are merely coincidence, or real cause-effect relationships needs further studying.

A molecular dissection, where DNA sequences in a genomic region are disrupted to test their effects on large-scale chromatin organization, epigenetic modifications and transcriptional activities, would greatly advance our understandings of the regulation of large-scale chromatin organization and its functional roles. However, such molecular dissection is difficult in mammalian systems. First, the mammalian genome has too many functionally redundant *cis*-elements and long-range interactions that disrupting one or several *cis*-elements might not produce a significant phenotype. Second, manipulating sequences in mammalian cells is still non-trivial. Finally, to exceed the resolution limit of light microscopy, the genomic region for dissection needs to be long enough, which would make the first problem worse.

Alternatively, using BAC transgene arrays as a model system could circumvent these difficulties. BACs contain 100-300 kb mammalian genomic inserts and can integrate into the mammalian genome as tandem arrays of up to Mb in size (Hu *et al.*, 2009; Sinclair *et al.*, 2010). These arrays have much lower DNA complexity than endogenous genomic regions of comparable size, yet they retain near normal levels of expression for the genes contained within the BAC genomic inserts (Hu *et al.*, 2009; Li *et al.*, 2009; Sinclair *et al.*, 2010). Moreover, BAC DNA sequences can be efficiently manipulated by BAC recombineering within *Escherichia coli* (*E. coli*).

Previously, our lab showed large scale chromatin structures of several BAC transgene arrays in several mammalian cell lines (Hu *et al.*, 2009; Bian and Belmont, 2010; Sinclair *et al.*, 2010), as well as BAC transgene motion after transcription activation (Khanna, Hu and Belmont, 2014; Kim *et al.*, 2019). We also dissected the HBB BAC for nuclear periphery targeting mechanisms (Bian *et al.*, 2013). Interestingly, however, a recent study in our lab showed that a reporter mini-gene had similar expression levels, uniformness of expression and long-term expression stability embedded in BACs containing transcriptionally active vs inactive genomic inserts (Chapter 2), indicating that the BAC transgenes could not recapitulate all features of the corresponding genomic regions, or the reporter mini-gene is insensitive to large-scale chromatin environments, or both.

Here we show distinctive chromatin structures formed by three BACs from that study (Chapter 2), demonstrating that BACs can reconstitute differential chromatin compaction, and that large-scale chromatin folding is regulated by DNA sequences. We also show that a reporter mini-gene could disrupt the nuclear periphery localization of the

HBB BAC without opening the condensed chromatin formed by the BAC transgene. Moreover, we show uncorrelated level of chromatin compaction and level of transcription from both reporter mini-gene and BAC transgens. These results suggest an experimental path to dissect the relationship between large-scale chromatin structure, nuclear localization, epigenetic modifications, and transcriptional activities using BAC transgene arrays.

RESULTS

Establishing BAC cell lines

Previously our lab analyzed the expression of a human UBC promoter driven, CpG free, GFP-Zeocin resistance fusion mini-gene (UBC-GFP-ZeoR) embedded in BACs containing transcriptionally active vs inactive genomic regions in mouse NIH 3T3 fibroblast cells (Chapter 2). Interestingly, this study found that the UBC-GFP-ZeoR reporter mini-gene had similar expression level, uniformness of expression and long-term expression stability in all the BACs tested (Chapter 2). To test whether BACs containing highly transcribed regions would form less condensed chromatin structure than BACs containing transcriptionally inactive regions, we chose three BACs from the previous study (Chapter 2) to study.

The DHFR-UG BAC was derived from the CITB-057L22 BAC (DHFR BAC) containing a mouse *Dhfr* gene and a partial mouse *Msh3* gene, both actively transcribed in NIH 3T3 cells (Figure 3.1a). The HBB-UG BAC was derived from the CTD-2643I7 BAC (HBB BAC) containing the transcriptionally silenced human β-globin locus (Figure

3.1b). The 2207K13-UG BAC was derived from the CTD-2207K13 BAC (2207K13 BAC) containing a human genomic region with no known genes or regulatory elements (Figure 3.1c). All three BACs have the UBC-GFP-ZeoR mini-gene inserted (Figure 3.1a-c). The UBC-GFP-ZeoR expression level/copy number in the three BACs was 1.9 in the DHFR-UG BAC, 1.5 in the HBB-UG BAC, and 1.2 in the 2207K13-UG BAC (Chapter 2).

To obtain more clones with large integrated BAC transgene arrays for easier chromatin structure analysis by light microscopy, we FACS sorted mixed clonal populations stably transfected with the BACs for high GFP expressing cells and isolated single clonal populations. As the previous study (Chapter 2) showed that in NIH 3T3 cells, a large fraction of stably transfected clones contained episomal BAC transgenes rather than integrated BAC transgenes, and that BAC episome clones had variegated GFP expression, we screened the clones for uniform GFP expression by flow cytometry.

We measured BAC copy number in the selected clones using qPCR (Table 3.1) and estimated BAC transgene array size by multiplying BAC copy number with BAC size (Table 3.1). We also calculated GFP fluorescence per BAC copy number and confirmed that the UBC-GFP-ZeoR expression is similar in all BAC clones (Table 3.1).

BAC transgene arrays reconstitute distinctive large-scale chromatin structures

We visualized the DHFR-UG, HBB-UG and 2207K13-UG BAC transgene arrays in the selected clones with 3D DNA FISH (Cremer *et al.*, 2008; Solovei and Cremer, 2010). Clones with more than one FISH spots per nucleus or with bad cell morphologies were excluded from the analysis. In general, the DHFR-UG BAC transgene arrays

formed an “open” chromatin structure, with elongated and fiber like FISH spots (Figure 3.2a). Moreover, with the increase of BAC transgene copy number, the DHFR-UG BAC transgene array FISH spots increased significantly in size, and became more and more fiber like (Figure 3.2a). Interestingly, the fiber like structure of the DHFR BAC transgene arrays, especially the large ones, were not uniform in diameter, but like a set of connected round or elongated blobs instead (Figure 3.2a). In contrast, both the HBB-UG and the 2207K13-UG BAC transgene arrays formed round blob like structures which became slightly irregular and increased slightly in size with the increase of BAC copy number (Figure 3.2b-c). There was no significant differences between the HBB-UG and the 2207K13-UG BAC transgene arrays (Figure 3.2b-c) as judged by eye.

To confirm the visual observations of the BAC transgene arrays, we quantitated the maximum intensity z-projections of the FISH spots by two metrics: area and circularity (defined as 4π×area/perimeter^2). A perfect circle has a circularity equal to 1 while a line has a circularity equal to 0. To objectively select the FISH spots, several segmentation methods were tested with a few FISH images and the accuracy was judged by eye. The multi-threshold maximum entropy thresholding method (Sacha, 1985) was selected for its accuracy and not requiring manual input. The whole analysis used Fiji (Schindelin *et al.*, 2012) and was summarized in Figure 3.3a. To separate BAC transgenes from background fluorescent aggregates and endogenous *Dhfr-Msh3* loci, areas containing individual nuclei were first cropped out from the z-projection images and the lowest thresholds produced by the 3-threshold maximum entropy thresholding function were used to further confine the areas containing the BAC transgene FISH spots. The middle thresholds were then used to segment the real FISH signals from the halos

surrounding the FISH signals (Figure 3.3b). The selected FISH spots were verified by eye and the area and circularity were measured. This analysis pipeline worked well with >90% of the FISH images analyzed (Figure 3.3b).

We plotted area or circularity against BAC transgene array size. Consistent with the visual observation, the area curve of the DHFR-UG BAC transgene array was always above that of the HBB-UG or 2207K13-UG BAC transgene arrays, while the circularity curve always below (Figure 3.3c). Moreover, with increasing BAC transgene array size, the area and the circularity of the DHFR-UG BAC transgene array changed much faster than the other two BAC transgene arrays (Figure 3.3c). Interestingly, the 2207K13-UG BAC transgene array had a slightly larger area than but similar circularity as the HBB-UG BAC transgene array. More sensitive imaging method, such as TEM, is needed to confirm whether 2207K13-UG and HBB-UG BAC transgenes can form different chromatin structures.

In conclusion, the DHFR-UG, HBB-UG and 2207K13-UG BACs can reconstitute differentially compacted chromatin in NIH 3T3 cells, indicating that the level of chromatin compaction is regulated by certain genomic sequences and that BAC transgene arrays are suitable for dissecting *cis*-elements regulating differential chromatin compaction.

The UBC-GFP-ZeoR reporter mini-gene moved the HBB BAC transgene away from the nuclear periphery

It is intriguing how the UBC-GFP-ZeoR mini-gene gets as highly expressed in a condensed chromatin as in an “open” chromatin. While we did not study the mechanism

of such transcriptional insensitivity to large scale chromatin structure here, we report an interesting discovery about the UBC-GFP-ZeoR reporter mini-gene and the HBB BAC.

A previous study in our lab showed that the HBB-LacO BAC (Figure 3.4a), derived by inserting a 10-kb, 256mer lac operator direct repeat (LacO) into the HBB BAC, localized at the nuclear periphery in NIH 3T3 cells (Bian *et al.*, 2013). Three functionally redundant nuclear periphery targeting regions (PTR1-3) on the HBB BAC (Figure 3.4a) were identified (Bian *et al.*, 2013). Deletion HBB BACs containing none of the PTRs, such as HBBD4, localized to chromocenters, which are clusters of pericentromeric heterochromatin, in NIH 3T3 cells, whereas deletion HBB BACs containing any of the PTRs, such as HBBD5 (Figure 3.4a), localized to the nuclear periphery (Bian *et al.*, 2013). Moreover, reducing H3K9me3 by knocking down SuvH1 and SuvH2 inhibited both the nuclear periphery and the chromocenter targeting of the HBB-LacO BAC transgene (Bian *et al.*, 2013).

Interestingly, the HBB-UG BAC has the UBC-GFP-ZeoR reporter mini-gene inserted in the PTR2 region (Figure 3.1b and Figure 3.4a). During our analysis of the HBB-UG BAC transgene chromatin structure, we found most of the DNA FISH spots were localized in the nuclear interior.

To validate this observation, we analyzed the nuclear localization of the HBB-LacO BAC transgene in clone HBB-LacO-C3, the HBBD4 BAC transgene in clone HBBD4-C40, the HBBD5 BAC transgene in clone HBBD5-C43, and the HBB-UG transgene in three clones: HBB-UG-fD2 (Chapter 2), HBB-UG-H3-50-4, and HBB-UG-H4-100-16. We also included two clones containing the DHFR-UG BAC, DHFR-UG-P4-14 and DHFR-UG-f3-15 as controls.

Although the clones from the previous study (Bian *et al.*, 2013) expressed an EGFP-dimer lac repressor-NLS (nuclear localization signal) fusion protein (EGFP-LacI), we visualized all the BAC transgenes by 3D DNA FISH. Optical sections where the FISH signals were in focus and was on the middle planes of the nuclei were used for analysis. FISH spots localized within 0.2 μm from the edge of the nuclei defined by DAPI staining was regarded as nuclear periphery localization. FISH spots completely or partially overlapping chromocenters was regarded as chromocenter localization.

Consistent with the previous study (Bian *et al.*, 2013), clone HBB-LacO-C3 and clone HBBD5-C43 had much higher rate of nuclear periphery localizing BAC transgenes than the two DHFR-UG BAC clones, whereas the HBBD4 clone had similar rate as the DHFR-UG BAC clones (Figure 3.4b). However, different form the previous study (Bian *et al.*, 2013), increased chromocenter localization was not observed in the HBBD4 clone and the summed rate of periphery localization and chromocenter localization of the HBBD4 clone was similar to that of the two DHFR-UG clones, and much lower than that of the HBB-LacO clone or that of the HBBD5 clone. This difference in chromocenter localization could be due to physiological changes in the clone, or due to possible distortions of chromocenters during the FISH procedure. The three HBB-UG clones had similar rate of nuclear periphery localizing and chromocenter localizing BAC transgenes as the two DHFR-UG clones.

In conclusion, insertion of the UBC-GFP-ZeoR mini-gene moved the HBB BAC transgene away from the nuclear periphery. However, whether the HBB-UG BAC transgene was moved to the nuclear interior or to the chromocenters needs further

confirmation, which would give a hint on whether H3K9me3 over the HBB BAC transgene was disrupted by the UBC-GFP-ZeoR mini-gene.

Two deletion DHFR BACs reconstituted differentially compacted chromatin

The observation that the UBC-GFP-ZeoR reporter gene expression was similar in "open" versus condensed chromatin formed by the DHFR-UG, HBB-UG and 2207K13-UG BACs indicates that at least for some genes the level of chromatin compaction is not directly related to the level of transcription. While here we do not dissect the underlying mechanism, we provide another example of uncorrelated transcription and chromatin compaction.

In an attempt to dissect the DHFR BAC for sequences responsible for the "open" chromatin structure formed by the BAC transgene array, we made two deletion DHFR BACs. The DHFR-c27d2 BAC contains a LacO, a CMV-mRFP-SV40-ZeoR expression cassette and a ~70 kb deletion of the 3' part of the *Msh3* gene (Figure 3.5a). The DHFR-c27d3crz BAC contains a LacO, a CMV-mRFP-SV40-ZeoR expression cassette and a ~80 kb deletion of the whole *Dhfr* gene and the 5' part of the *Msh3* gene (Figure 3.5a). An NIH 3T3 clone expressing an EGFP-LacI was used for making the BAC cell lines. We measured BAC copy number in individual clones by qPCR (Table 3.2) and estimated BAC transgene array size (Table 3.2).

The BAC transgene arrays were visualized directly in formaldehyde fixed cells. A previous study (Sinclair *et al.*, 2010) showed that in two NIH 3T3 clones stably transfected with a full length DHFR BAC containing LacO, the EGFP-LacI spots frequently formed extended, fiber-like conformations, with 6-7 BACs per EGFP-LacI

spot, and BAC and LacO DNA FISH signals were highly co-localized. Here we found that in a DHFR-c27d2 BAC clone (DHFR-c27d2-02) with high BAC copy number, the EGFP-LacI spots were tiny, well separated from each other, and arranged in fiber-like conformation (Figure 3.5b), similar to what was observed with the full-length DHFR BAC (Sinclair *et al.*, 2010). In contrast, a DHFR-c27d3crz BAC clone (DHFR-c27d3crz-65) with similar BAC copy number, the EGFP-LacI spots were frequently fused together, forming blob-like structures (Figure 3.5c). Such difference was less obvious in clones with low BAC copy numbers (DHFR-c27d2-17 vs DHFR-c27d3crz-33). However, we did find that clone DHFR-c27d2-17 had more EGFP-LacI spots than clone DHFR-c27d3crz-33 (Figure 3.5 b-c) in general. However, the DHFR-c27d3crz transgene arrays looked less condensed than the HBB-UG and 2207K13-UG BAC transgenes.

To validate the observed differences in the DHFR-c27d2 and the DHFRc27d3crz BAC transgene chromatin structures, we quantitated the maximum intensity z-projections of the EGFP-LacI spots. The whole analysis used Fiji (Schindelin *et al.*, 2012) and was summarized in Figure 3.6a. To separate BAC transgenes from background fluorescent aggregates, areas containing individual nuclei were first cropped out from the z-projection images and a maximum entropy thresholding method was used to further confine the areas containing the EGFP-LacI spots. We quantitated the EGFP-LacI spots in two aspects. First, the number, area and circularity of individual spots, segmented by an unsharp mask followed by maximum entropy or simple thresholding, were measured. Second, the contour of all the spots was identified by blurring the spots with a Gaussian mask followed by maximum entropy or simple thresholding, and the area and circularity of the contour were measured.

Consistent with visual observation, the two DHFR-c27d2 clones had more EGFP-LacI spots, with each spot having larger circularity and smaller area comparing to the two DHFR-c27d3crz clones. Moreover, DHFR-c27d2 clones had 9-16 BACs per EGFP-LacI spot (Table 3.2), which is around twice of the full length DHFR-c27 BAC (Sinclair *et al.*, 2010), while the DHFR-c27d3crz clones had 23-45 BACs per EGFP-LacI spot (Table 3.2). Considering that the DHFR-c27d2 BAC is half in size of the full length DHFR BAC, the compaction level of the DHFR-c27d2 should be similar to that of the full DHFR BAC.

Higher transcription level over the more condensed deletion DHFR BAC transgene array than the more "open" one

As the DHFR-c27d2 BAC retains the divergent promoter driving the *Dhfr* and the *Msh3* gene, whereas the DHFR-c27d3crz BAC does not contain any known promoters, we hypothesized that the divergent promoter and active transcription of the *Dhfr* and the *Msh3* gene could be contributing to the more "open" chromatin structure of the DHFR-c27d2 BAC transgene. Therefore, we measured transcription levels over the two deletion DHFR BACs and the full length DHFR BAC by qPCR, using cells without the BAC transgenes as reference samples. The transcription levels were further normalized by BAC copy number for easier comparison. As expected, the transcription levels of the DHFR-c27d2 BAC was similar to the corresponding regions on the full length DHFR BAC (Figure 3.7a). Surprisingly, the DHFR-c27d3crz BAC had even higher transcription than the full length DHFR BAC, in spite of the absence of the divergent promoter (Figure 3.7a), indicating alternative promoters getting activated by the deletion of the divergent promoter. Interestingly, the CMV-mRFP mini-gene embedded in the deletion DHFR

BACs had higher expression in the more condensed DHFR-c27d3crz BAC than in the more "open" DHFR-c27d2 BAC, as shown both by qPCR (Figure 3.7b) and by flow cytometry (Table 3.2). These results suggests that the differential chromatin compaction of the two deletion DHFR BACs is not directly related to level of transcription, and the expression of the CMV-mRFP mini-gene embedded in the BACs is not purely determined by large-scale chromatin structure.

DISCUSSION

With accumulating studies showing correlations between large-scale chromatin organization and genome function, dissecting the determinants regulating large-scale chromatin organization becomes more and more important. However, the high density of regulatory elements and the frequent long-range interactions of the mammalian genome makes molecular dissection difficult. Here we show that BAC transgene arrays could be a powerful model system for dissecting the mechanisms regulating large-scale chromatin organization.

Differential chromatin compaction has long been observed. Here we show that we can reproducibly reconstitute distinctive large-scale chromatin conformations with BAC transgene arrays by altering the genomic inserts contained within the BACs. Moreover, we show that we can manipulate the chromatin conformation of the BAC transgenes by manipulating the BAC sequences.

It is not clear what the "default" chromatin conformation is, where neither transcriptional activation or repression is present. A previous study (Boettiger *et al.*, 2016)

has shown that genomic regions with active histone modifications have more "open" chromatin conformation than genomic regions with repressive histone modifications, and genomic regions with neither active or repressive histone modifications have chromatin conformation in between. While we have not analyzed the epigenetic marks over the BAC transgenes, genomic mapping studies in fibroblasts have shown that the genomic regions corresponding to the DHFR BAC has elevated active histone modifications, that corresponding to the HBB BAC has elevated H3K9me3, and that corresponding to the 2207K13 BAC is low in both active and repressive histone modifications. Consistent with the previous study (Boettiger *et al.*, 2016), the DHFR-UG BAC has the most "open" and the HBB-UG BAC most condensed large-scale chromatin conformation. Interestingly, the 2207K13-UG BAC forms similar chromatin conformation as the HBB-UG BAC. Further studies comparing the epigenetic modifications over the BAC transgenes would yield helpful information about the relationship between epigenetic modifications and large-scale chromatin structures.

Position effects and position effects variegation have been commonly seen in transgene expression (Robertson *et al.*, 1995; Gierman *et al.*, 2007; Ramunas *et al.*, 2007; Akhtar *et al.*, 2013; Tchasovnikarova *et al.*, 2015). A previous study (Gierman *et al.*, 2007) has shown that a human PGK promoter driven GFP reporter gene integrated at different chromosomal positions has expression levels corresponding to the transcriptional activity of the regions of integration, and that transcriptionally active regions of integration have a more open chromatin structure than inactive regions. Interestingly, however, here we show that the UBC-GFP-ZeoR had similar expression levels in the DHFR-UG, HBB-UG and 2207K13-UG BACs, despite of the distinctive

large-scale chromatin structures and expression levels of the BAC transgenes, indicating position-independent expression of the UBC-GFP-ZeoR reporter gene. However, a previous study has shown that the UBC-GFP-ZeoR integrated into the chromosomes alone without the BACs has copy-number independent, position dependent expression (Chapter 2). A possible explanation is that the UBC-GFP-ZeoR reporter gene itself is insensitive to the chromatin environment of the integration site, but multi-copy plasmid repeats induced silencing in certain genomic regions. Alternatively, the BACs we selected happen to not contain the real repressive genomic regions, or the BAC transgenes cannot recapitulate certain features required for repressing reporter gene expression. Further studying the mechanism enabling the high expressing of the UBC-GFP-ZeoR reporter gene in condensed BAC transgene arrays would reveal interesting clues about transcriptional regulation and large-scale chromatin structures.

MATERIALS AND METHODS

BAC modifications

The DHFR-UG, HBB-UG and 2207K13-UG BACs were constructed in another study (Chapter 2). The DHFR-UG BAC was derived from the CITB-057L22 BAC (DHFR BAC) containing mouse chr13:92,992,156-93,161,185 (mm9). The HBB-UG BAC was derived from the CTD-2643I7 BAC (HBB BAC) containing human chr11:5,218,904-5,426,232 (hg19). The 2207K13-UG BAC was derived from the CTD-2207K13 BAC (2207K13 BAC), containing human chr1:79,180,278-79,286,094. The UBC-GFP-ZeoR mini-gene, derived from plasmid pUGG (Chaturvedi *et al.*, 2018), was

inserted into these BACs at the following positions in relative to the mouse or human genome: chr13:93,099,101-93,099,102 (mm9) in the DHFR BAC, chr11:5,390,233-5,390,244 (hg19) in the HBB BAC, and chr1:79,224,725-79,224,726 (hg19) in the 2207K13 BAC.

Construction of the two deletion DHFR BACs, DHFR-c27d2 and DHFR-c27d3crz was described in another study (Chapter 2). The DHFR-c27 BAC (50) containing a 256-mer lac operator direct repeat (LacO) and a CMV-mRFP-SV40-ZeoR expression cassette was derived from the DHFR BAC. The DHFR-c27d2 and the DHFR-c27d3crz BAC were derived from the DHFR-c27 BAC. The DHFR-c27d2 BAC contains a ~70 kb deletion of the 3' part of the *Msh3* gene (chr13:92,992,394-93,059,992, mm9). The DHFR-c27d3crz BAC contains a ~80 kb deletion of the whole *Dhfr* gene and the 5' part of the *Msh3* gene (chr13:93,075,639-93,160,310, mm9), and a new CMV-mRFP-SV40-ZeoR expression cassette introduced at the remaining part of *Msh3* gene (chr13:93,007,742-93,007,743, mm9).

Cell culture and BAC cell line establishment

NIH 3T3 cells were cultured with Dulbecco's modified Eagle medium (DMEM, with 4.5 g/L D-glucose, 4 mM L-glutamine, 1 mM sodium pyruvate and 3.7 g/L NaHCO3) supplemented with 10% HyClone Bovine Growth Serum (GE Healthcare Life Sciences, Cat. # SH30541.03).

Mixed clonal populations stably transfected with the DHFR-UG, HBB-UG or 2207K13-UG BAC, respectively, obtained from another study (Chapter 2) were first FACS sorted for high GFP expressing cells as described in section "Flow Cytometry".

The cells were either directly sorted into 96-well plates, at 1 cell/well, or sorted into tubes. The cells sorted into tubes were cultured for several days for them to recover and expand, and then diluted and plated in a 96-well plate at ~1 cell per well to obtain single populations. Both the mixed clonal and single clonal populations with the DHFR-UG, HBB-UG or 2207K13-UG BAC were maintained with 75 μg/ml Zeocin (Thermo Fisher Scientific).

NIH 3T3 cell clone 3T3_LG_C29 stably expressing an EGFP-dimer lac repressor-NLS (nuclear localization signal) fusion protein (EGFP-LacI) was obtained from another study (Bian *et al.*, 2013) and was maintained with 200 μg/ml Hygromycin B (Thermo Fisher Scientific). Clone 3T3_LG_C29 was transfected with the DHFR-c27d2 or the DHFR-c27d3crz BAC using Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's instructions and selected with 200 μg/ml Hygromycin B and 75 μg/ml Zeocin for 1-2 weeks. Individual cell clones were obtained by serial dilution (Strukov and Belmont, 2008). Cells stably transfected with the DHFR-c27d2 or the DHFR-c27d3crz BAC were maintained with 200 μg/ml Hygromycin B and 75 μg/ml Zeocin.

NIH 3T3 clones DHFR-UG-f3-1 and DHFR-UG-f3-15, stably transfected with the DHFR-UG BAC, and NIH 3T3 clone HBB-UG-fD2 stably transfected with the HBB-UG BAC were created in another study (Chapter 2). 3T3_LG_C29 derived clones, HBB-LacO-C3, HBBD4-C40 and HBBD5-C43 were created in another study (Bian *et al.*, 2013). NIH 3T3 clone DHFR-c27-4 stably transfected with the DHFR-c27 BAC was created in a previous study (Bian and Belmont, 2010).

Flow Cytometry

Flow cytometry used similar protocols as another study (Chapter 2). GFP expression from the UBC-GFP-ZeoR mini-gene was analyzed on a BD FACS Canto II Flow Cytometry Analyzer (BD Biosciences), using the FITC/Alexa Fluor-488 channel (488nm laser, 502 longpass dichroic mirror and 530/30 bandpass filter). mRFP expression from the CMV-mRFP-SV40-ZeoR mini-gene was analyzed on a BD LSR Fortessa (BD Biosciences), using the PE channel (561 nm laser and 582/15 nm bandpass filter). Rainbow fluorescent beads (Spherotech, Cat. # RFP-30-5A) were used as fluorescence intensity standards.

Cell sorting used a BD FACS AriaII (BD Biosciences) and was assisted by the flow cytometry facility stuff at Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign (UIUC). Two sorting windows were set on the FITC channel (488 nm laser, 505 longpass dichroic mirror and 530/30 nm bandpass filter), one over the highest GFP signals and one near the right half of the GFP peak, each took up ~10% of parent cells.

Estimation of BAC transgene copy number

BAC transgene copy number within individual cell clones was measured by real-time quantitative PCR (qPCR), using relative quantitation methods as described in another study (Chapter 2). Mouse genes *Sgk1* and *Hprt1* were used as endogenous controls, assuming four copies of each gene per cell in NIH 3T3. Multiple primer pairs were used for estimating the copy numbers of the BAC transgenes: five for the DHFR-UG BAC, seven for the HBB-UG BAC, six for the 2207K13-UG BAC, and three for

DHFR-c27d2 and DHFR-c27d3crz. ΔC_T method (Equation 3.1 and 3.2) was used to calculate the copy numbers of the PCR amplification regions on the UBC-GFP-ZeoR reporter gene, on the HBB BAC or on the 2207K13 BAC, and $\Delta\Delta C_T$ method (Equation 3.3 and 3.4) was used to calculate the copy numbers of the PCR amplification regions on the DHFR BAC. The mean copy number of all PCR amplification regions was calculated as the copy number of that region. All primers are listed in Data A.1.

$$\Delta C_T = C_{T\,\text{test region}} - (C_{T\,Sgk1} + C_{T\,Hprt1})/2 \quad 3.1$$

$$\text{copy number}_{\Delta C_T} = 4 \times 1.95^{-\Delta C_T} \quad 3.2$$

$$\Delta\Delta C_T = \Delta C_{T\,\text{transgene clone}} - \Delta C_{T\,\text{NIH 3T3}} \quad 3.3$$

$$\text{copy number}_{\Delta\Delta C_T} = 4 \times 1.95^{-\Delta\Delta C_T} \quad 3.4$$

3D DNA FISH

Biotin labeled DNA FISH probes were made from BAC DNA by end-labeling with Terminal Transferase, according to a published protocol (Dernburg, 2011). The following reagents were used: AluI, DpnI, HaeIII, MseI, MspI, RsaI (New England Biolabs) and CutSmart Buffer (comes with the enzymes); Terminal Deoxynucleotidyl Transferase and reaction buffer (Thermo Fisher Scientific, Cat. # EP0161); dATP (New England Biolabs, Cat. # N0446S) and Biotin-14-dATP (Thermo Fisher Scientific, Cat. # 19524016).

DNA FISH of interphase nuclei was performed according to published protocols (Cremer *et al.*, 2008; Solovei and Cremer, 2010) with small modifications, and was described in another study (Chapter 2). FISH signals were detected by incubation with

Alex Fluor 488 conjugated Strepavidin (1:500; Jackson ImmunoResearch, Cat. # 016-540-084), Alexa Fluor 647 conjugated Strepavidin (1:200; Jackson ImmunoResearch, Cat. # 016-600-084) or Alexa 594 conjugated Strepavidin (1:200; Life Technology, Cat. # S11227).

Microscopy

For examining EGFP-LacI signals, cells were grown on coverslips and fixed with 3-4% paraformaldehyde in DPBS before mounting. All samples- including fixed cells expressing EGFP-LacI and 3D DNA FISH sample- were mounted with a Mowiol-DABCO anti-fade medium ('Mowiol mounting medium', 2006) containing ~3 μg/ml DAPI (MilliporeSigma). 3D z-stack images were acquired using a V4 OMX (GE healthcare) microscope, equipped with a 100X, 1.4 NA oil immersion objective (Olympus), and two Evolve EMCCDs (Photometrics). Images were deconvolved using the deconvolution algorithm (Agard *et al.*, 1989) provided by the *softWoRx* software (GE Healthcare). Chromatic aberrations were measured using the alignment slide provided by GE Healthcare and the OMX Image registration function in the *softWoRx* was used to correct the chromatic aberrations in all DNA FISH images according to the manufacturer's instructions.

Analysis of DNA FISH images

The area and circularity of z-projected DNA FISH signals were measured using Fiji (Schindelin *et al.*, 2012). A macro was developed to semi-automate the analysis process. First, a maximum intensity z-projection of a 3D z-stack image was created using

the “Z Project…” function with “projection=[Max Intensity]”. Next, individual nucleus was cropped out from the z-projection image to minimize the detection of background fluorescent aggregates by the computer program. An “Otsu” auto-thresholding function was applied to the DAPI channel of the z-projection to select individual nuclei. The selection was turned into a rectangle and enlarged by 1 μm. Next, a “Maximum Entropy Multi-Threshold” function with “number=3” from the IJ Plugins package (http://ij-plugins.sourceforge.net/index.html) was applied to the FISH channel of the cropped nucleus image, generating three thresholds. To prevent endogenous *Dhfr-Msh3* loci and background fluorescent aggregates from being identified, the area containing the FISH signal was first identified by applying the “Analyze Particles...” function with “size=0-Infinity display exclude clear add" to the FISH channel masked with the lowest threshold and selecting the largest particle. The FISH signal was then identified by masking the area containing the FISH signal with the middle threshold. The selection of FISH signal was examined manually and any area not corresponding to the BAC transgene arrays were deselected. Finally, the area and circularity of the correctly selected FISH signals were measured by the “Set Measurements...” function with “area perimeter shape redirect=None decimal=3” and the “Measure” function.

The following clones were used for Figure 3.3c: DHFR-UG-f3-1, -f3-15, -P4-14; HBB-UG-fD2, -H3-50-4, -H4-100-16; 2207K13-UG-K3-50-17, -K4-100-12.

Analysis of BAC transgene nuclear localization

The BAC transgenes were visualized by DNA FISH and images were analyzed using Fiji (Schindelin *et al.*, 2012). The orthogonal view of the z-stack images were

examined manually and optical sections where he FISH spots were both in focus and were at the middle planes of the nuclei were used for analysis. To detect nuclear periphery localization, the edges of the nuclei was selected by applying a "Gaussian Blur…" function with "sigma=2.50" and subsequently a "Default dark" auto-thresholding function to the DAPI channel. The selection was then shrank by 0.2 μm. FISH spots overlapping the selection were regarded as localized at the nuclear periphery. FISH spots completely or partially overlapping a chromocenter were regarded as localized at chromocenters.

95% confidence intervals of the proportions were calculated by Equation 3.5 and 3.6 (N- total number of nuclei; N_d- number of nuclei with periphery or chromocenter localized BAC transgenes; p_U and p_L- upper and lower limits of the 95% confidence interval, respectively).

$$\sum_{k=0}^{N_d} \binom{N}{k} \cdot p_U^k (1-p_U)^{N-k} = 0.05/2 \tag{3.5}$$

$$\sum_{k=0}^{N_d-1} \binom{N}{k} \cdot p_L^k (1-p_L)^{N-k} = 1 - 0.05/2 \tag{3.6}$$

Analysis of EGFP-LacI images

The conformation and arrangement of z-projected EGFP-LacI spots were measured using Fiji (Schindelin *et al.*, 2012). Macros were developed to semi-automate the analysis process. To prevent the computer program from identifying background fluorescent aggregates as EGFP-LacI spots, 100 pixel × 100 pixel areas containing the EGFP-LacI spots were first identified. First, maximum intensity z-projections of 3D z-

stack images were created and individual nucleus were cropped out from the z-projection images, using the same procedures as described in section “Analysis of DNA FISH images”. Incorrectly segmented individual nucleus images were then removed from analysis manually. Next, 100 pixel × 100 pixel areas containing EGFP-LacI spots were cropped out from the individual nucleus images, by applying a “MaxEntropy dark” auto-thresholding function and subsequently a “Analyze Particles...” function with “size=20-Infinity pixel exclude clear add” to the FITC channel, followed by turning the selection bounds of all identified particles into 100 pixel × 100 pixel square selections. Incorrectly segmented EGFP-LacI images were then removed from analysis manually.

To make a selection contouring all of the EGFP-LacI spots, a “Gaussian Blur...” function with “sigma=2.5” was applied to the FITC channel, followed by a “MaxEntropy dark” auto-thresholding function (clone DHFR-c27d2-17 and clone DHFR-c27d3crz-33) or by setting threshold at 25% of the maximum intensity (clone DHFR-c27d2-02 and clone DHFR-c27d3crz-65). To segment individual EGFP-LacI spots, an “Unsharp Mask...” function with “radius=1 mask=0.90” was applied to the FITC channel, followed by a “MaxEntropy dark” auto-thresholding function (clone DHFR-c27d2-17 and clone DHFR-c27d3crz-33) or by setting threshold at 25% of the maximum intensity (clone DHFR-c27d2-02 and clone DHFR-c27d3crz-65). The area and circularity of the selections were measured by “Set Measurements...” with “area mean standard modal min centroid center perimeter bounding shape limit add redirect=None decimal=3” and “Analyze Particles...” with “show=Overlay display exclude include”. Finally, for clone DHFR-c27d2-17 and clone DHFR-c27d3crz-33, each segmented images were examined manually and any incorrectly selected background spots were removed from the results.

The number of individual spots in each nucleus were calculated by counting the number of entries for the individual spots for the nucleus.

Estimation of transcription level of the full length DHFR BAC and the two deletion DHFR BACs

Transcription levels over the DHFR-c27d2 and the DHFR-c27d3crz BAC were measured by real-time quantitative PCR (qPCR), using relative quantitation. Cellular RNA was extracted using RNeasy Mini Kit (QIAGEN) with in-solution DNase digestions, according to the manufacturer's instructions. RNA was reverse transcribed with random hexamers using qScript Flex cDNA Kit (Qaunta Biosciences). Mouse gene *Hprt1* were used as endogenous control. Multiple primer pairs were used for measuring the transcription over the BAC transgenes. ΔC_T method (Equation 3.7 and 3.8) was used to calculate the transcription level of the CMV-mRFP reporter gene, and $\Delta\Delta C_T$ method (Equation 3.9 and 3.10) was used to calculate the transcription level of regions on the DHFR BAC. All primers are listed in Data A.1.

$$\Delta C_T = C_{T\,\text{test region}} - C_{T\,Hprt1} \tag{3.7}$$

$$\text{transcription level}_{\Delta C_T} = 4 \times 1.95^{-\Delta C_T} \tag{3.8}$$

$$\Delta\Delta C_T = \Delta C_{T\,\text{transgene clone}} - \Delta C_{T\,\text{3T3\_LG\_C29}} \tag{3.9}$$

$$\text{transcription level}_{\Delta\Delta C_T} = 4 \times 1.95^{-\Delta\Delta C_T} \tag{3.10}$$

Transcription level over the DHFR-c27 BAC was measured similar as the deletion DHFR BACs, except that the *Actb* gene was used as endogenous control and

NIH 3T3 cells were used as the reference sample instead of clone 3T3_LG_C29 in Equation 3.7.

Code availability

All computational scripts used for image analysis are available at https://bitbucket.org/Binhui/image_analysis/src.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of General Medical Sciences R01 GM058460 to A.S.B. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

FIGURES AND TABLES

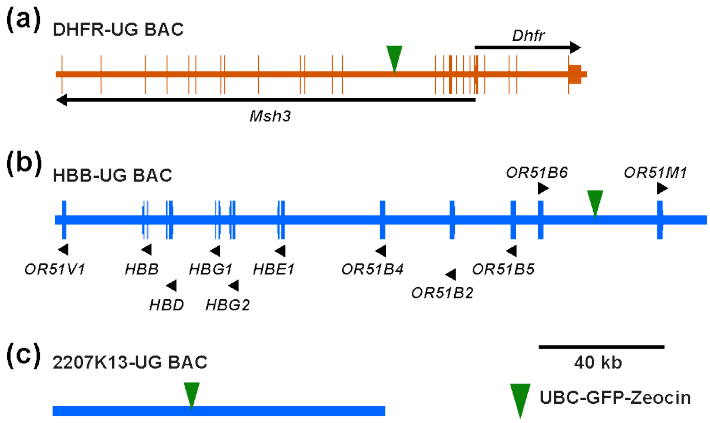

Figure 3.1. Maps of the DHFR-UG, HBB-UG and 2207K13-UG BAC. Longer vertical bars- exons; shorter vertical bars- UTRs; black arrows or arrowheads- direction of transcription; green arrow heads- UBC-GFP-ZeoR insertion site.

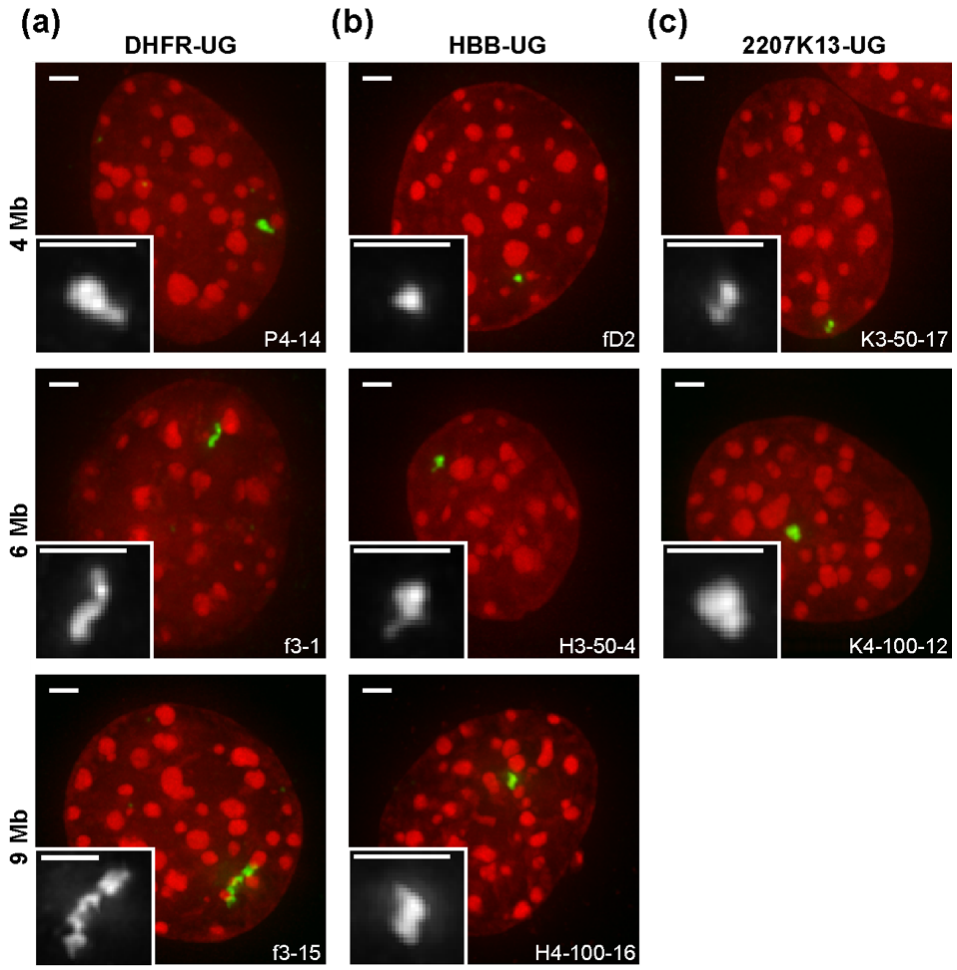

Figure 3.2. Representative 3D DNA FISH images showing distinctive chromatin conformation formed by the DHFR-UG, HBB-UG and 2207K13-UG clones. Maximum intensity z-projections of three DHFR-UG clones (a), three HBB-UG clones (b), and two 2207K13-UG clones (c) are shown. Insets are enlarged FISH signals. Clone numbers are shown on the bottom right of each image. Estimated sizes of the BAC transgene arrays are shown on the left. Red- DAPI; Green- FISH signal; Scale bar = 2 μm.

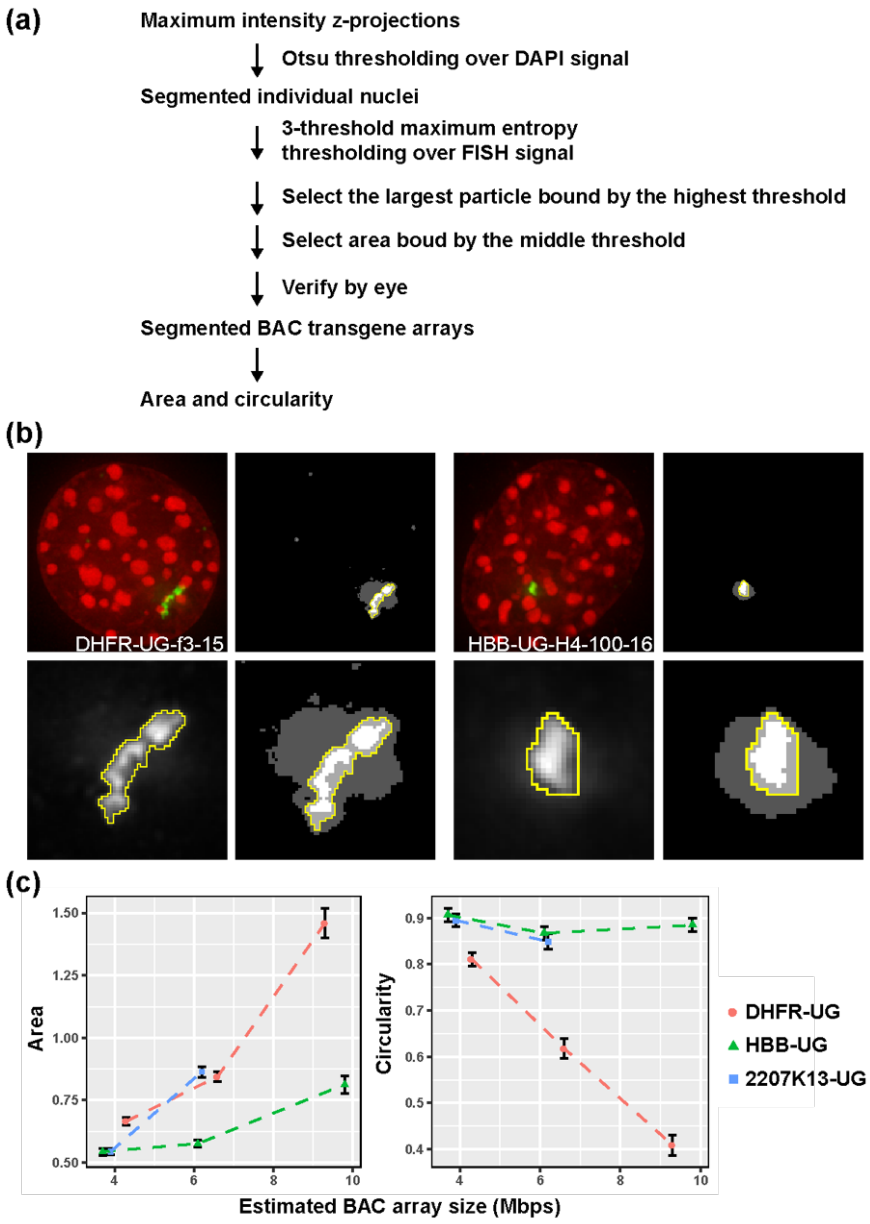

Figure 3.3. Quantitative analysis of the FISH images confirms differences in large-scale chromatin folding. (a) Work flow of the analysis. (b) Example images showing the results of segmentation of FISH signals in a DHFR-UG clone and a HBB-UG clone. (c) Median area (left, y-axis, unit = μm2) or median circularity (right, y-axis) of the FISH signals are plotted against estimated BAC transgene array size (Mb). Error bars represent standard error; N = 56 ~81.

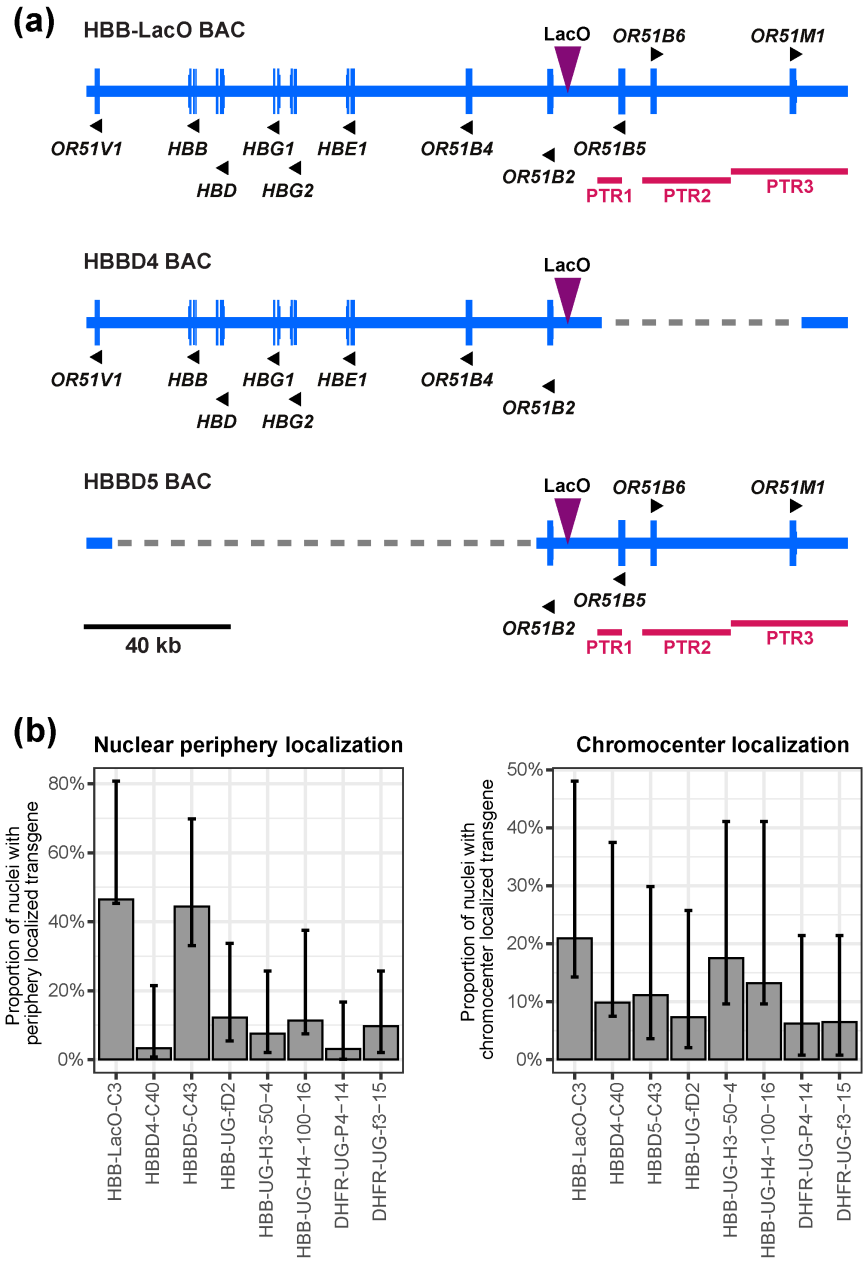

Figure 3.4. The UBC-GFP-ZeoR reporter mini-gene moved the HBB BAC transgene away from the nuclear periphery. (a) Maps of the HBB-LacO, HBBD4 and HBBD5 BACs. Dashed lines- deleted regions; PTR1-3: periphery targeting regions identified in another study (Bian *et al.*, 2013); Longer vertical bars- exons; shorter vertical bars- UTRs; black arrows or arrowheads- direction of transcription; purple arrow heads- LacO insertion site. (b) Rate of nuclear periphery localization (left) and chromocenter

Figure 3.4. Cont.

localization (right) of the BAC transgenes in the corresponding clones. Error bars represent 95% confidence intervals. Number of images analyzed: 43 for HBB-LacO-C3, 61 for HBBD4-C40, 36 for HBBD5-C43, 41 for HBB-UG-fD2, 40 for HBB-UG-H3-50-4, 53 for HBB-UG-H4-100-16, 32 for DHFRUG-P4-14, 31 for DHFR-UG-f3-15.

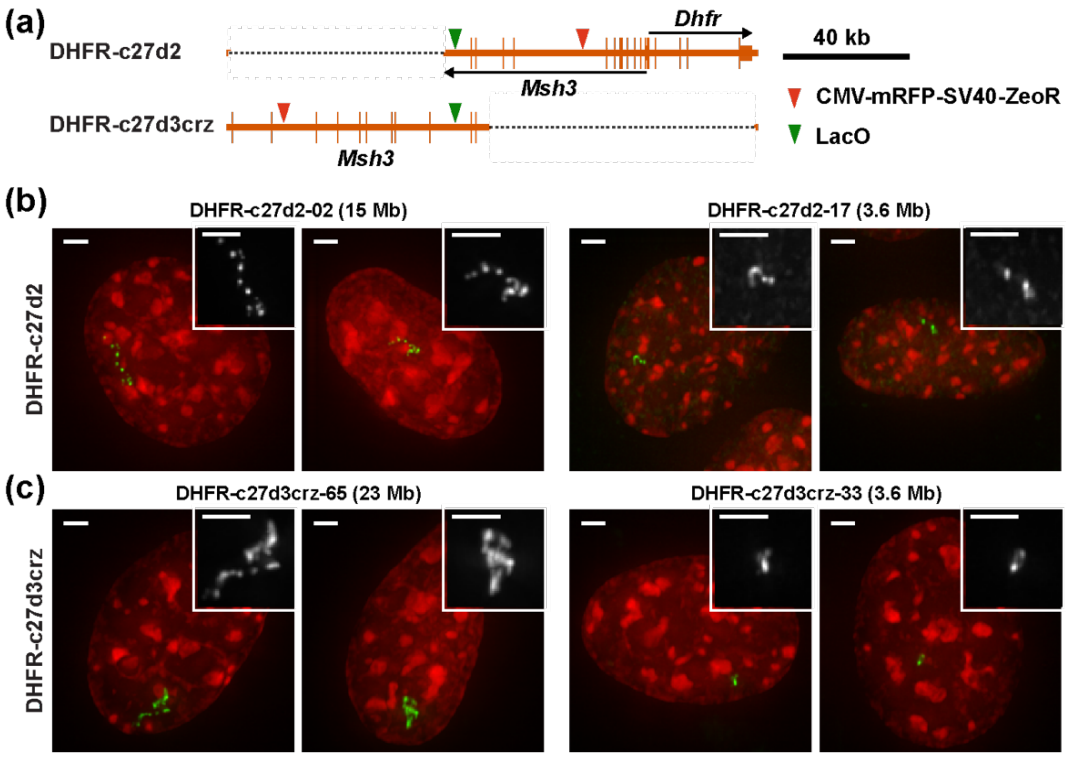

Figure 3.5. The DHFR-c27d2 deletion DHFR BAC formed more "open" chromatin conformation than the DHFR-c27d3crz deletion DHFR BAC. (a) Maps of the DHFR-c27d2 and the DHFR-c27d3crz BACs. Dashed lines- deleted regions; Longer vertical bars- exons; shorter vertical bars- UTRs; black arrows or arrowheads- direction of transcription; green arrow heads- LacO insertion site; red arrow heads- CMV-mRFP-SV40-ZeoR expression cassette insertion site; (b) Representative images of a large insertion and a small insertion DHFR-c27d2 clones. (c) Representative images of a large insertion and a small insertion DHFR-c27d3crz clones. (b-c) Maximum intensity z-projections are shown. Insets are enlarged EGFP-LacI spots. Estimated sizes of the BAC transgene arrays are behind the clone names. Red- DAPI; Green- EGFP-LacI; Scale bars = 2 μm.

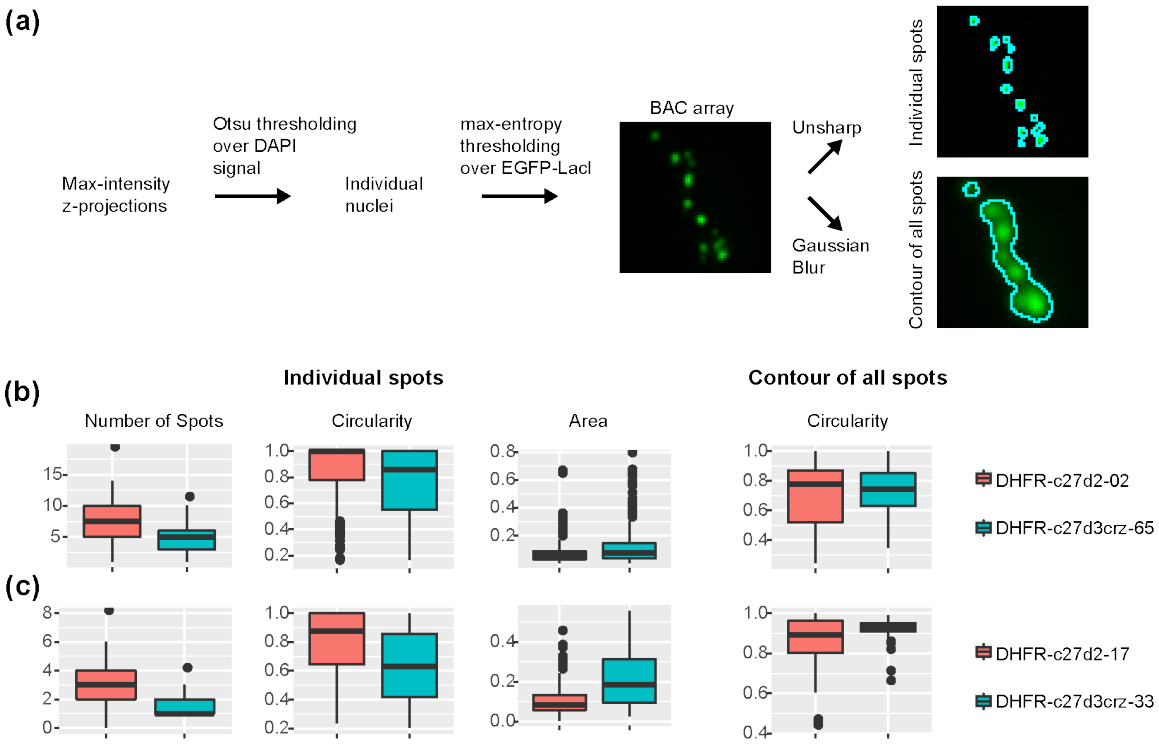

Figure 3.6. Quantitative analysis of the EGFP-LacI images confirms differences in large-scale chromatin folding. (a) Work flow of the analysis. (b) Comparison of two large insertion clones, DHFR-c27d2-02 vs DHFR-c27d3crz-65. (c) Comparison of two small insertion clones, DHFR-c27d2-17 vs DHFR-c27d3crz-33.

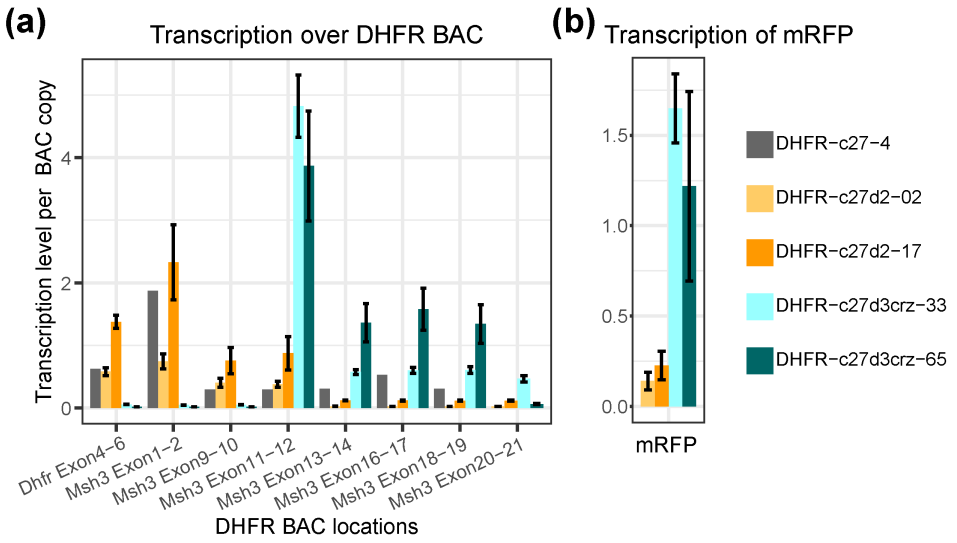

Figure 3.7. Transcription over the DHFR-c27d3crz BAC is higher than DHFR-c27d2 BAC. (a) Transcription level over the DHFR BAC normalized by BAC copy number in two DHFR-27d2 clones and two DHFR-27d3crz clones, with comparison of clone DHFR-c27-4. (b) Transcription level of the mRFP reporter gene in the same clones as in (a). x-axis- locations on the DHFR BAC; y-axis- transcription level normalized by copy number; Error bars represent standard deviations of three independent experiments.

Table 3.1. Information of the DHFR-UG, HBB-UG and 2207K13-UG clones.

BAC	clone #	BAC copy #	BAC size (kb)	BAC transgene array size (Mb)	GFP/beads	GFP/beads/BAC copy #	GFP CV%
DHFR-UG	f3-1	37	179	6.6	94.3	2.5	43.2
DHFR-UG	f3-15	52	179	9.3	80.6	1.6	46.2
DHFR-UG	P4-14	24	179	4.3	52.1	2.2	26.6
HBB-UG	fD2	17	218	3.7	30.0	1.8	38
HBB-UG	H3-50-4	28	218	6.1	54.5	1.9	33.4
HBB-UG	H4-100-16	45	218	9.8	120.4	2.7	37.7
2207K13-UG	K3-50-17	33	116	3.9	/	/	/
2207K13-UG	K4-100-12	53	116	6.2	112.7	2.1	30.8

Table 3.2. Information of the DHFR-c27d2 and DHFR-c27d3crz clones.

BAC	clone #	BAC copy #	BAC size (kb)	BAC transgene array size (Mb)	mRFP/ beads	mRFP/beads/ BAC copy #	mRFP CV%	EGFP-LacI spot #	BAC copy #/EGFP-LacI spot
DHFR-c27d2	02	123	124	15.1	4.39	0.036	112.8	7.7	16.0
DHFR-c27d2	17	29	124	3.6	1.36	0.047	194.9	3.1	9.4
DHFR-c27d3crz	65	216	107	23.1	17.69	0.082	123	4.8	45.0
DHFR-c27d3crz	33	34	107	3.6	7.75	0.228	114.8	1.5	22.7

REFERENCES

Agard, D. A. *et al.* (1989) 'Fluorescence microscopy in three dimensions.', *Methods in cell biology*, 30, pp. 353–77.

Akhtar, W. *et al.* (2013) 'Chromatin position effects assayed by thousands of reporters integrated in parallel.', *Cell*. Elsevier, 154(4), pp. 914–27.

Bazett-Jones, D. P. and Hendzel, M. J. (1999) 'Electron spectroscopic imaging of chromatin.', *Methods (San Diego, Calif.)*, 17(2), pp. 188–200.

Belmont, A. S. *et al.* (1989) 'Large-scale chromatin structural domains within mitotic and interphase chromosomes in vivo and in vitro.', *Chromosoma*, 98(2), pp. 129–43.

Bian, Q. *et al.* (2013) 'β-Globin cis-elements determine differential nuclear targeting through epigenetic modifications.', *The Journal of cell biology*, 203(5), pp. 767–83.

Bian, Q. and Belmont, A. S. (2010) 'BAC TG-EMBED: one-step method for high-level, copy-number-dependent, position-independent transgene expression.', *Nucleic acids research*, 38(11), p. e127.

Bian, Q. and Belmont, A. S. (2012) 'Revisiting higher-order and large-scale chromatin organization.', *Current opinion in cell biology*. Elsevier Ltd, 24(3), pp. 359–66.

Boettiger, A. N. *et al.* (2016) 'Super-resolution imaging reveals distinct chromatin folding for different epigenetic states.', *Nature*. Nature Publishing Group, 529(7586), pp. 418–22.

Bohrmann, B. and Kellenberger, E. (1994) 'Immunostaining of DNA in electron microscopy: an amplification and staining procedure for thin sections as alternative to gold labeling.', *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, 42(5), pp. 635–43.

Chambeyron, S. and Bickmore, W. a (2004) 'Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription.', *Genes & development*, 18(10), pp. 1119–30.

Chaturvedi, P. *et al.* (2018) 'Stable and reproducible transgene expression independent of proliferative or differentiated state using BAC TG-EMBED.', *Gene therapy*, 25(5), pp. 376–391.

Chen, Y. *et al.* (2018) 'Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler', *Journal of Cell Biology*, 217(11), pp. 4025–4048.

Cremer, M. *et al.* (2008) 'Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes.', *Methods in molecular biology (Clifton, N.J.)*, 463, pp. 205–39.

Cremer, T. and Cremer, C. (2001) ‘Chromosome territories, nuclear architecture and gene regulation in mammalian cells.’, *Nature reviews. Genetics*, 2(4), pp. 292–301.

Dernburg, A. F. (2011) ‘Fragmentation and labeling of probe DNA for whole-mount FISH in Drosophila.’, *Cold Spring Harbor protocols*, 2011(12), pp. 1527–30.

Dixon, J. R. *et al.* (2012) ‘Topological domains in mammalian genomes identified by analysis of chromatin interactions.’, *Nature*. Nature Publishing Group, 485(7398), pp. 376–80.

Dixon, J. R., Gorkin, D. U. and Ren, B. (2016) ‘Chromatin Domains: The Unit of Chromosome Organization’, *Molecular Cell*. Elsevier Inc., 62(5), pp. 668–680.

Fawcett, D. W. (1966) ‘On the occurrence of a fibrous lamina on the inner aspect of the nuclear envelope in certain cells of vertebrates’, *American Journal of Anatomy*, 119(1), pp. 129–145.

Ghirlando, R. and Felsenfeld, G. (2013) ‘Chromatin structure outside and inside the nucleus.’, *Biopolymers*, 99(4), pp. 225–32.

Gierman, H. J. *et al.* (2007) ‘Domain-wide regulation of gene expression in the human genome’, *Genome Research*, 17(9), pp. 1286–1295.

Guelen, L. *et al.* (2008) ‘Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.’, *Nature*, 453(7197), pp. 948–51.

Hu, Y. *et al.* (2009) ‘Large-scale chromatin structure of inducible genes: transcription on a condensed, linear template’, *The Journal of Cell Biology*, 185(1), pp. 87–100.

Khanna, N., Hu, Y. and Belmont, A. S. S. (2014) ‘HSP70 Transgene Directed Motion to Nuclear Speckles Facilitates Heat Shock Activation’, *Current Biology*. Elsevier Ltd, 24(10), pp. 1138–1144.

Kim, J. *et al.* (2019) ‘Nuclear speckle fusion via long-range directional motion regulates speckle morphology after transcriptional inhibition’, *Journal of Cell Science*, (March), p. jcs.226563.

van Koningsbruggen, S. *et al.* (2010) ‘High-Resolution Whole-Genome Sequencing Reveals That Specific Chromatin Domains from Most Human Chromosomes Associate with Nucleoli’, *Molecular Biology of the Cell*. Edited by A. G. Matera, 21(21), pp. 3735–3748.

Li, Y. *et al.* (2009) ‘Mutant LRRK2(R1441G) BAC transgenic mice recapitulate cardinal features of Parkinson’s disease.’, *Nature neuroscience*, 12(7), pp. 826–8.

Lund, E. *et al.* (2013) ‘Lamin A/C-promoter interactions specify chromatin state-dependent transcription outcomes.’, *Genome research*, 23(10), pp. 1580–9.

‘Mowiol mounting medium’ (2006) *Cold Spring Harbor Protocols*, 2006(1), p. pdb.rec10255.

Németh, A. *et al.* (2010) ‘Initial Genomics of the Human Nucleolus’, *PLoS Genetics*. Edited by A. Akhtar, 6(3), p. e1000889.

Olins, A. L. and Olins, D. E. (1974) ‘Spheroid chromatin units (v bodies).’, *Science (New York, N.Y.)*, 183(4122), pp. 330–2.

Ou, H. D. *et al.* (2017) ‘ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells’, *Science*, 357(6349).

Parada, L. A. *et al.* (2002) ‘Conservation of relative chromosome positioning in normal and cancer cells.’, *Current biology : CB*, 12(19), pp. 1692–7.

Peric-Hupkes, D. *et al.* (2010) ‘Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.’, *Molecular cell*, 38(4), pp. 603–13.

Pickersgill, H. *et al.* (2006) ‘Characterization of the Drosophila melanogaster genome at the nuclear lamina.’, *Nature genetics*, 38(9), pp. 1005–14.

Quinodoz, S. A. *et al.* (2018) ‘Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus’, *Cell*. Elsevier Inc., 174(3), p. 744–757.e24.

Ramunas, J. *et al.* (2007) ‘Real-time fluorescence tracking of dynamic transgene variegation in stem cells.’, *Molecular therapy : the journal of the American Society of Gene Therapy*, 15(4), pp. 810–7.

Robertson, G. *et al.* (1995) ‘Position-dependent variegation of globin transgene expression in mice.’, *Proceedings of the National Academy of Sciences of the United States of America*, 92(12), pp. 5371–5.

Robson, M. I. *et al.* (2016) ‘Tissue-Specific Gene Repositioning by Muscle Nuclear Membrane Proteins Enhances Repression of Critical Developmental Genes during Myogenesis’, *Molecular Cell*, 62(6), pp. 834–847.

Sacha, J. (1985) ‘Maximum Entropy Thresholding’, (i), p. 1985.

Schindelin, J. *et al.* (2012) ‘Fiji: an open-source platform for biological-image analysis.’, *Nature methods*, 9(7), pp. 676–82.

Schöfer, C. and Weipoltshammer, K. (2018) ‘Nucleolus and chromatin.’, *Histochemistry and cell biology*, 150(3), pp. 209–225.

Shopland, L. S. *et al.* (2003) ‘Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: Evidence for local euchromatic neighborhoods’, *Journal of Cell Biology*, 162(6), pp. 981–990.

Sinclair, P. *et al.* (2010) ‘Dynamic plasticity of large-scale chromatin structure revealed by self-assembly of engineered chromosome regions.’, *The Journal of cell biology*, 190(5), pp. 761–76.

Solovei, I. and Cremer, M. (2010) ‘3D-FISH on cultured cells combined with immunostaining.’, *Methods in molecular biology (Clifton, N.J.)*, 659, pp. 117–26.

Strukov, Y. G. and Belmont, a. S. (2008) ‘Development of Mammalian Cell Lines with lac Operator-Tagged Chromosomes’, *Cold Spring Harbor Protocols*, 2008(2), p. pdb.prot4903-pdb.prot4903.

Tchasovnikarova, I. A. *et al.* (2015) ‘Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells’, *Science*, 348(6242), pp. 1481–1485.

Wang, Q. *et al.* (2016) ‘Cajal bodies are linked to genome conformation’, *Nature Communications*. Nature Publishing Group, 7, pp. 1–17.

Xu, J. *et al.* (2018) ‘Super-Resolution Imaging of Higher-Order Chromatin Structures at Different Epigenomic States in Single Mammalian Cells’, *Cell Reports*. ElsevierCompany., 24(4), pp. 873–882.

APPENDIX A: SUPPLEMENTARY DATA FOR CHAPTER 2

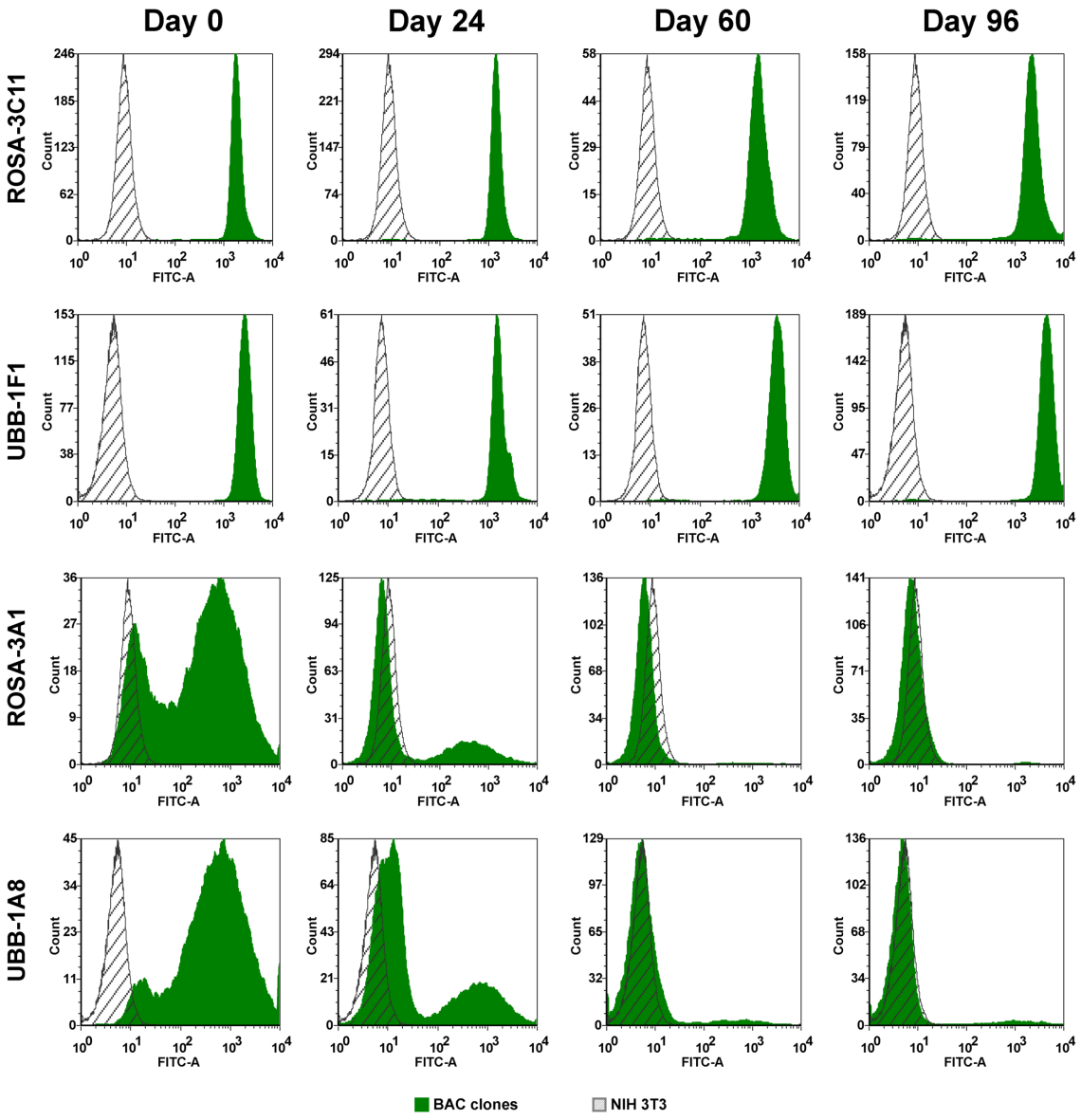

Figure A.1. GFP fluorescence histogram of representative "uniform" and "heterogeneous" expressing NIH 3T3 clones at day 0, 24, 60 and 96 without selection obtained by flow-cytometry. Gray- autofluorescence of untransfected cells; Green- GFP fluorescence of the indicated clones; x-axis- fluorescence; y-axis- cell number.

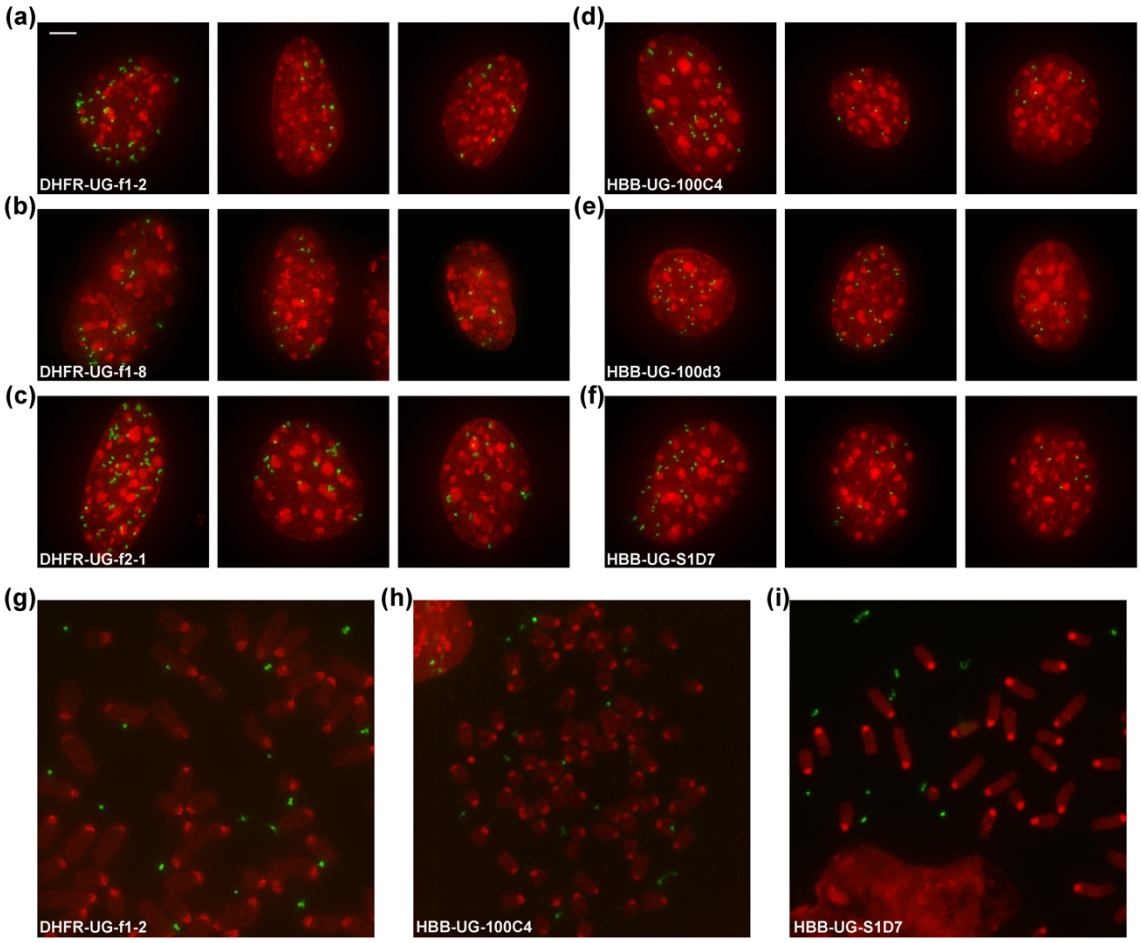

Figure A.2. DNA FISH over heterogeneously expressing clones transfected with DHFR-UG or HBB-UG BAC. (a-f) 3D DNA FISH over interphase nuclei from three DHFR-UG BAC (a-c) and three HBB-UG BAC (d-f) heterogeneous clones using BAC probes. Maximum-intensity projections are shown. (g-i) DNA FISH over mitotic spreads of one DHFR heterogeneous clone (g) and two HBB heterogeneous clones (h-i) using BAC probes. Red- DNA DAPI stain; Green- BAC FISH signal. Scale bar = 4 μm.

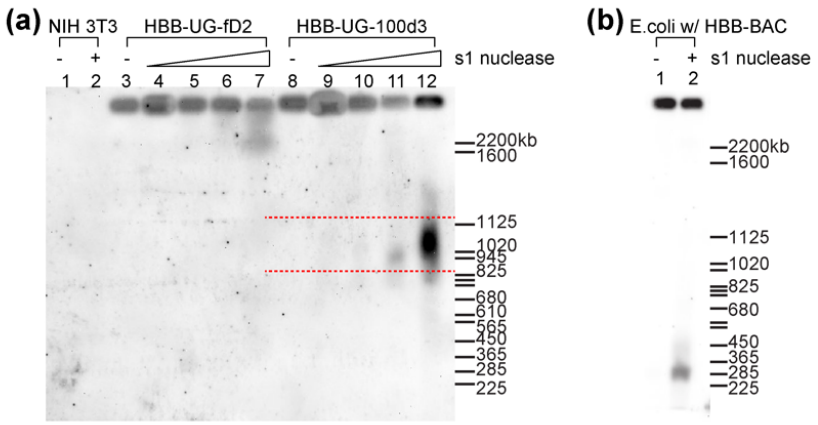

Figure A.3. Southern blot hybridization - using probes prepared from BAC DNA- of cellular DNA without enzyme digestion, or digested with increasing amount of S1 Nuclease, separated by PFGE. (a) Lane 1-2: NIH 3T3 cellular DNA; Lane 3-7: uniform clone HBB-UG-fD2; Lane 8-12: heterogenous clone HBB-UG-100d3; (b) E. coli carrying the HBB BAC.

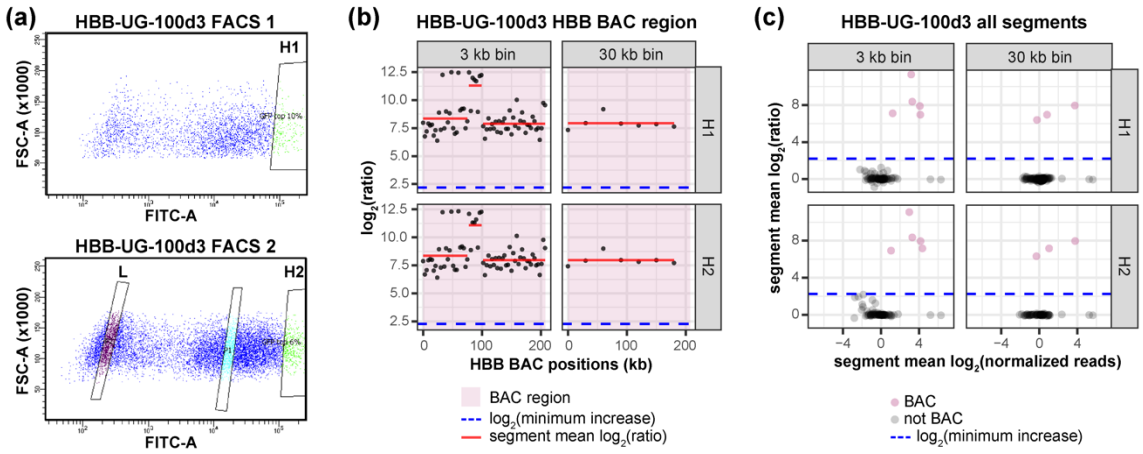

Figure A.4. CNV analysis of the HBB-UG-100d3 clone. (a) Two FACS experiments for collecting cells with high (H1 and H2), and low (L) fluorescence subpopulations. x-axis- FITC channel intensity; y-axis- forward scatter; H1, H2, and L- sorting windows. (b) $\log_2$(ratio) of individual bins (dark gray dots) and the segment mean $\log_2$(ratio) (red lines) over the HBB BAC (pink highlight) in the H1 and H2 subpopulations of the HBB-UG-100d3 clone. (c) Scatter plot of segment mean $\log_2$(ratio) vs segment mean $\log_2$(normalized reads) of all segments of the H1 and H2 subpopulations of the HBB-UG-100d3 clone. Pink dots- segments belonging to the HBB BAC, including the HBB locus, UBC-GFP-ZeoR and the BAC vector; Black dots- genomic segments. (b-c) Blue dashed line: $\log_2$(estimated minimum copy number increase).

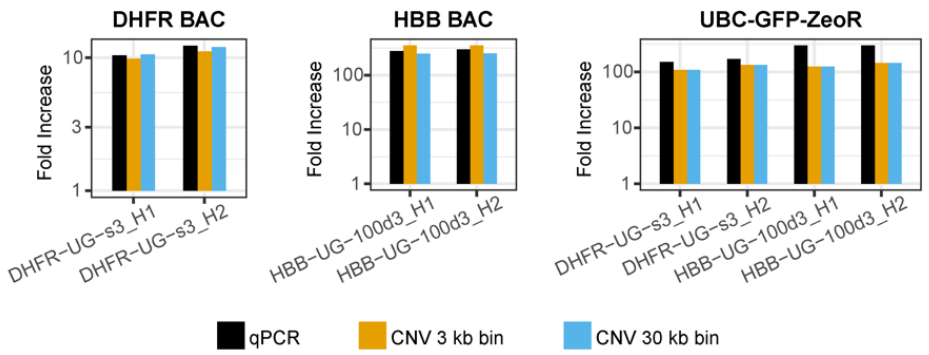

Figure A.5. Comparison of copy number fold increases of BAC regions (y-axis) in H1 and H2 relative to L subpopulations measured by CNV analysis versus by qPCR.

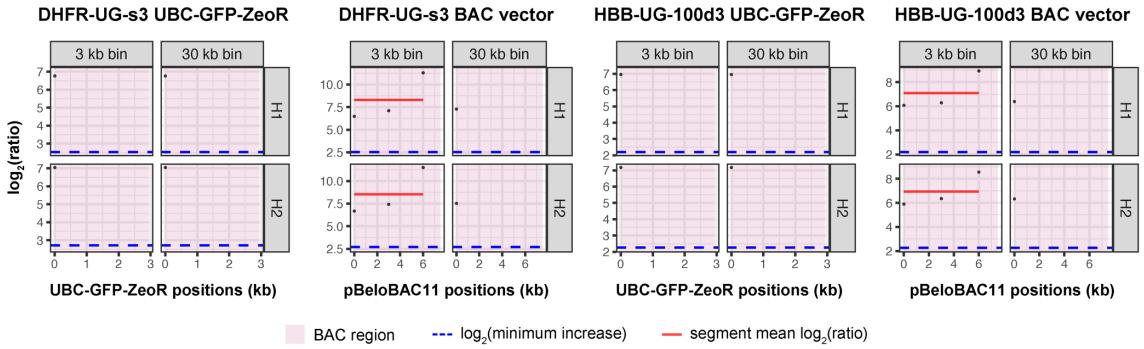

Figure A.6. $\log_2$(ratio) of individual bins (dark gray dots) and the segment mean $\log_2$(ratio) (red lines) over the UBC-GFP-ZeoR and the pBeloBAC11 BAC backbone, in the H1 and H2 subpopulations of the DHFR-UG-s3 and the HBB-UG-100d3 clones. Blue dashed line: $\log_2$(estimated minimum copy number increase).

(a)

Reporter Construct	Heterogeneous Clones%	Relative Promoter Strength
CMV–mRFP–SV40–ZeoR	0 (n=13)	1.0
B2M–mRFP–SV40–ZeoR	0 (n=6)	0.3
UBC–GFP–ZeoR	60% (n=30)	3.0
UBC–mRFP–SV40–ZeoR	45% (n=20)	3.0
RPL32–mRFP–SV40–ZeoR	35% (n=29)	5.5
EEF1α–mRFP–SV40–ZeoR	67% (n=18)	4.6

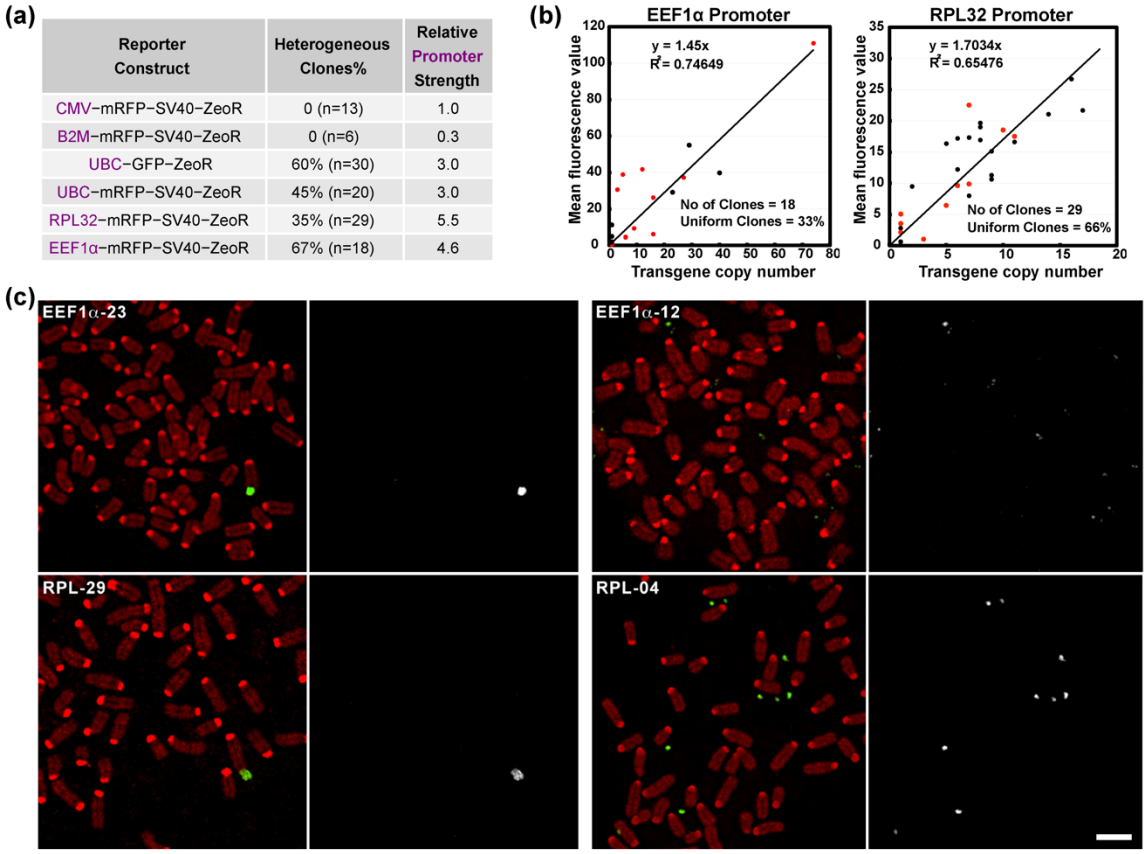

Figure A.7. Promoter strength and BAC episome formation. (a) Summary of percentages of heterogeneously expressing clones and relative promoter strengths for different reporter constructs embedded in the DHFR BAC. (b) Average normalized RFP fluorescence of individual cell clones (y-axis) are plotted versus transgene copy number (x-axis) for EEF1α-mRFP-SV40-ZeoR (left) or RPL32-mRFP-SV40-ZeoR (right) reporter constructs embedded in the DHFR BAC. Linear regression fits (black line) with y-intercepts set at 0 are shown with corresponding R-squared values and equations. Red circles- heterogeneous clones; Black circles- uniform clones. Bottom right: Number of clones analyzed and percentage of uniform clones. (c) Chromosomal locations of BAC transgene arrays within metaphase chromosomes of indicated clones as visualized by FISH using DHFR BAC probe (green) and DAPI staining (red). Scale bar = 5 μm.

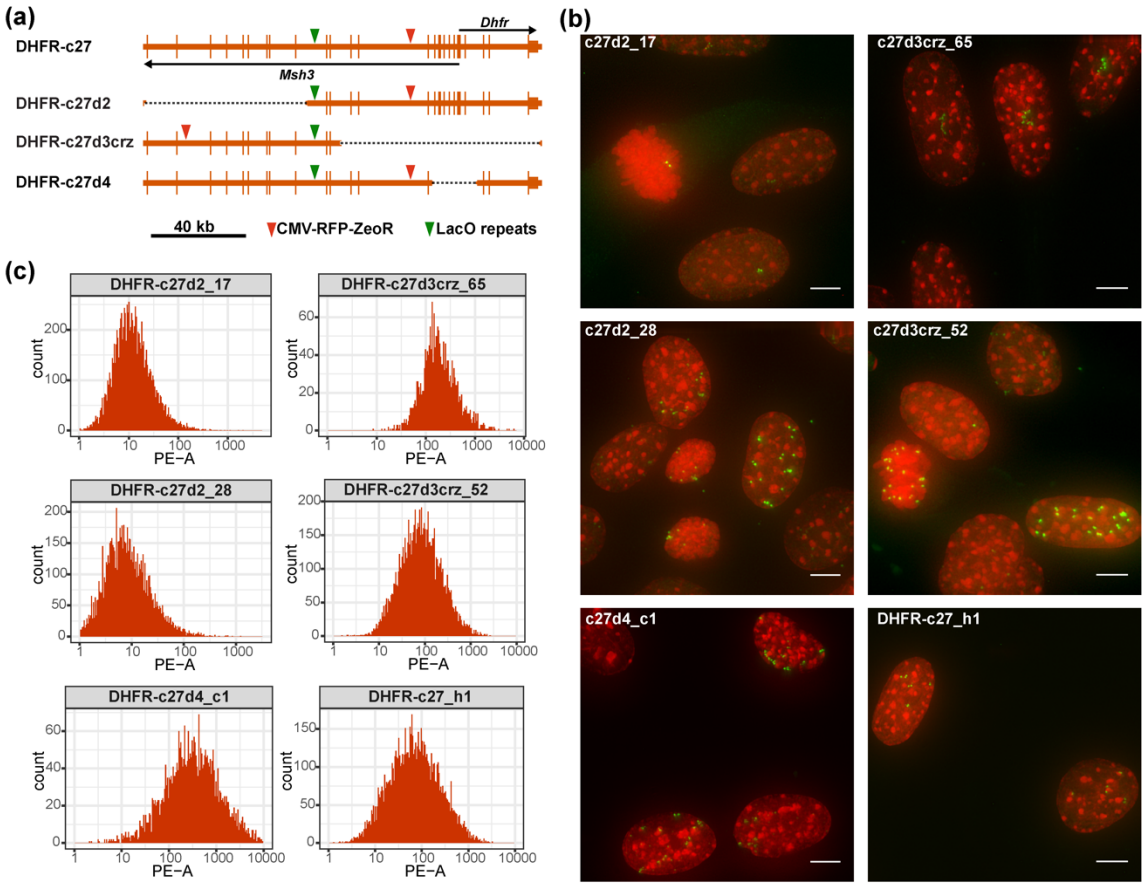

Figure A.8. Episome formation of DHFR BACs with CMV-mRFP-SV40-ZeoR reporter gene insertions. (a) Schematics of the intact DHFR BAC (DHFR-c27) and three DHFR BAC deletions (DHFR-c27d2, -c27d3crz and -c27d4). Longer vertical bars- exons; shorter vertical bars- UTRs; arrows- direction of transcription; black dashed lines- deleted regions; green arrowhead- Lac operator repeats (LacO) insertion site; red arrowhead- CMV-mRFP-SV40-ZeoR insertion site. (b) Representative maximum-intensity projection images of clones with integrated BACs (c27d2_17, c27d3crz_65) and clones with episomal BACs (remaining clones). Red- DNA DAPI staining; Green- EGFP-LacI. Gamma = 0.5 was applied to the green channels of all images, and to the red channels of clone c27d2-17 and c27d3crz-52 images after projection. Scale bars = 5 μm. (c) mRFP fluorescence histogram of the clones in (b) obtained by flow-cytometry. x-axis- signal from PE channel; y-axis- cell number.

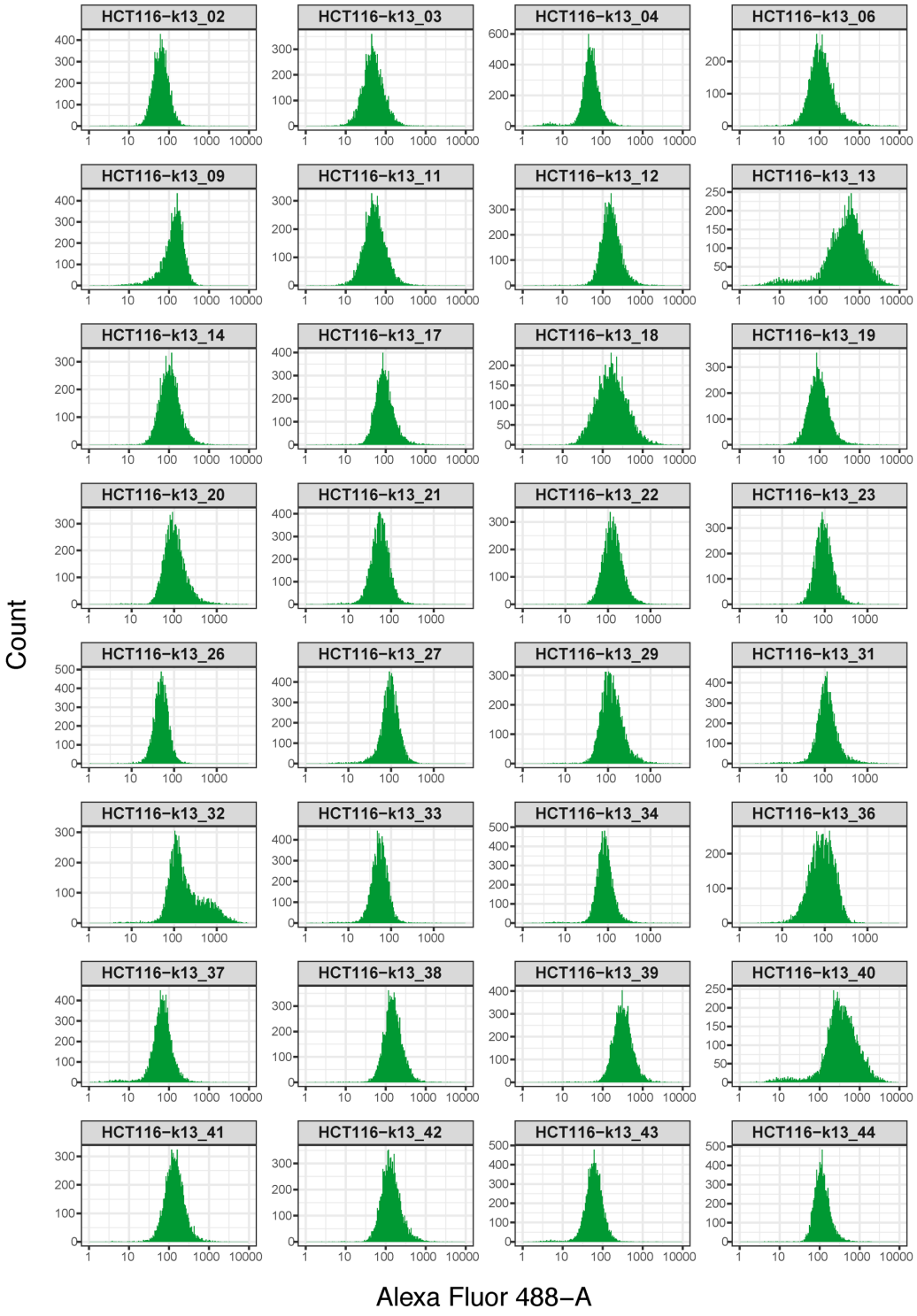

Figure A.9. GFP fluorescence histograms of HCT116 derived clones stably transfected with the 2207K13-UG BAC obtained by flow-cytometry. x-axis- signal from PE channel; y-axis- cell number.

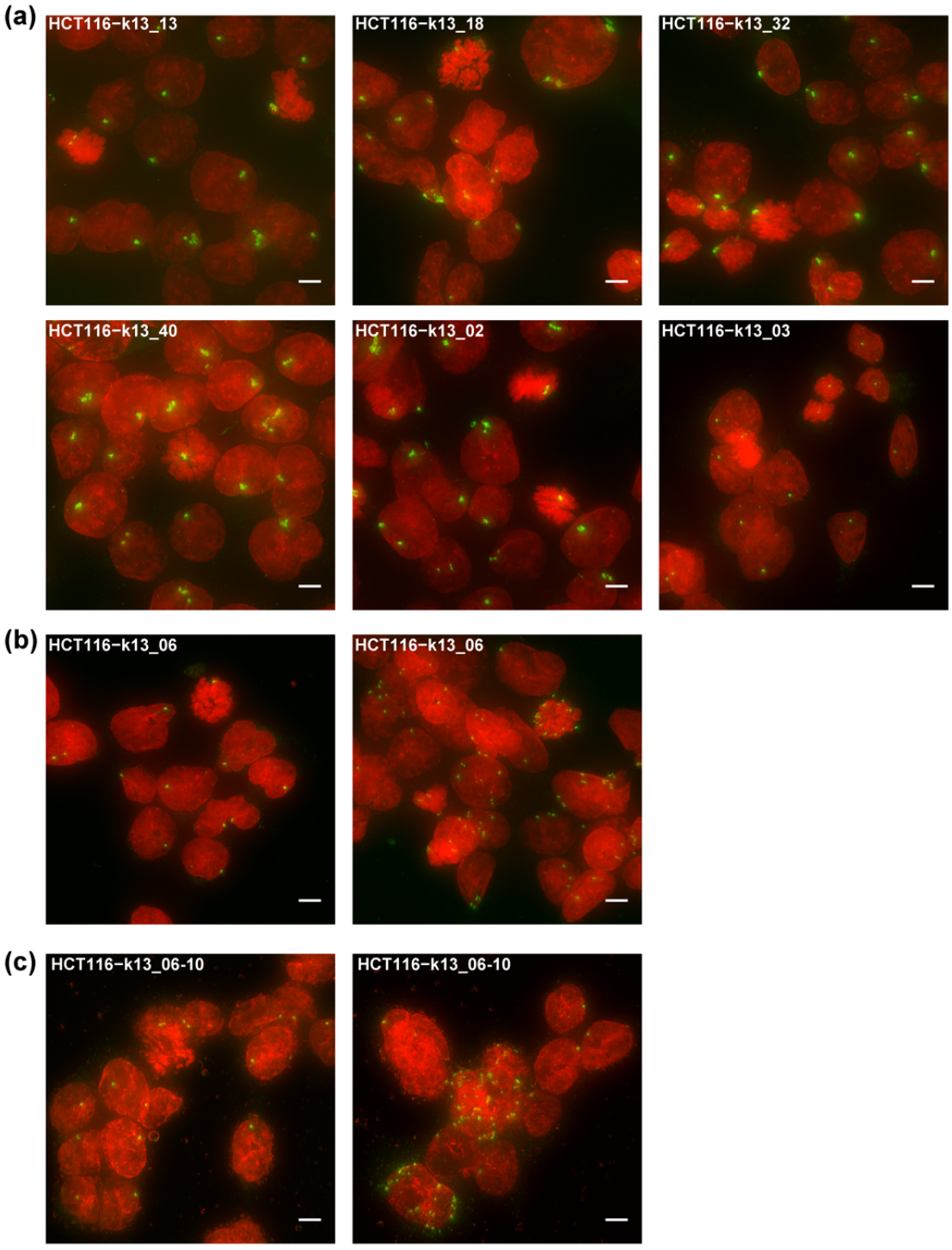

Figure A.10. 3D DNA FISH over HCT116 derived clones stably transfected with the 2207K13-UG BAC using BAC probes. (a) Four clones with broad GFP-reporter fluorescence histograms (Figure A.9) show integrated BAC arrays (top 3 panels, left

Figure A.10. Cont.

bottom panel), similar to clones with narrow GFP-reporter fluorescence histograms (middle and right bottom panels). (b-c) One clone with a narrow GFP-reporter fluorescence histogram showed a subpopulation of cells within the clonal population showing episomal BAC transgenes (b). Subcloning this clone identified a subclones which also contained a similar mixture of clones with integrated versus episomal forms of the the BAC transgenes (c). (a-c): Maximum intensity projections are shown. Gamma = 0.5 was applied to the green channels after projection. Red- DAPI staining; Green- FISH. Scale bars = 5 μm.

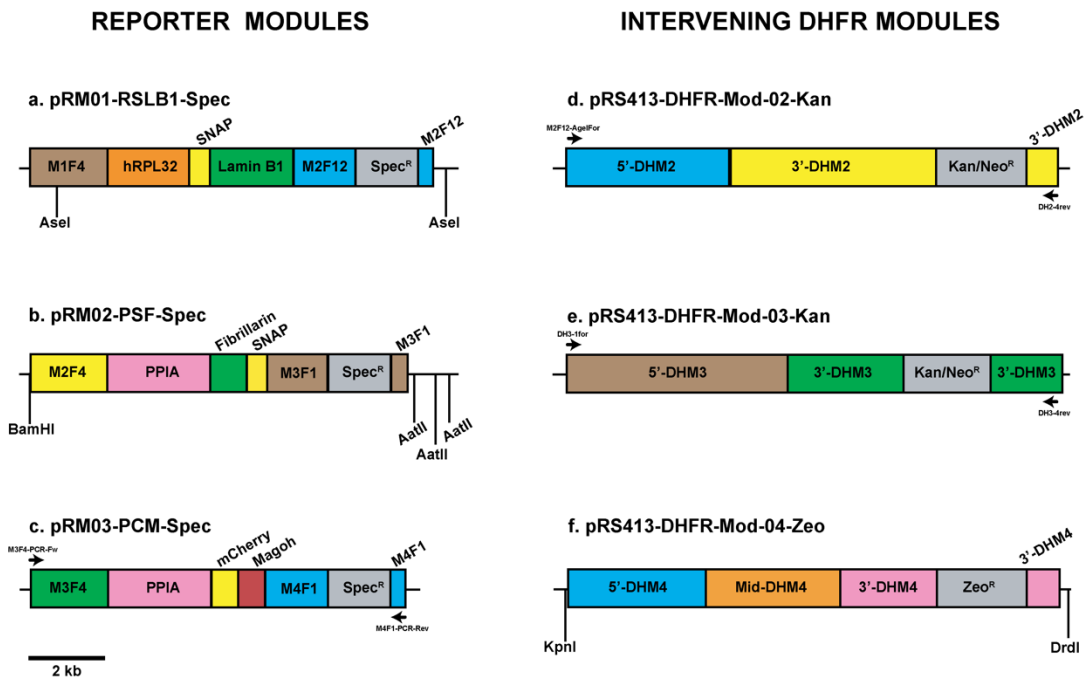

Figure A.11. Maps of Reporter and DHFR modules used for BAC-MAGIC. (a-c) reporter expression cassettes subcloned in the respective reporter recipient modules harboring SpecR selection marker (gray). (d-e) Schematics of the intervening DHFR modules harboring Kan/NeoR or ZeoR selection markers (gray). (a-e) Longer vertical bars represent the indicated restriction endonucleases used to generate recombineering fragments and arrows show the binding sites of primers used for amplification of recombineering fragments. See Methods for details of terminal regions corresponding to DHFR BAC homology regions. Scale bar = 2 kb.

Table A.1. Synthetic DNA fragment "RCS" sequence.

Name	Sequence	Notes
RCS	5' GGCCGCGGCGCGCCTTAATTAACCGGTG 3' 3' CGCCGCGCGGAATTAATTGGCCACGATC 5'	NotI overhang NheI overhang

Data A.1. List of primers and oligos used for Chapter 2 and Chapter 3. Primer/oligo name, sequence and description of all primers/oligos saved in an excel file (Data_A.1.xlsx).

APPENDIX B: BAC TRANSGENES DO NOT FORM EPISOMES IN MOUSE ES CELLS

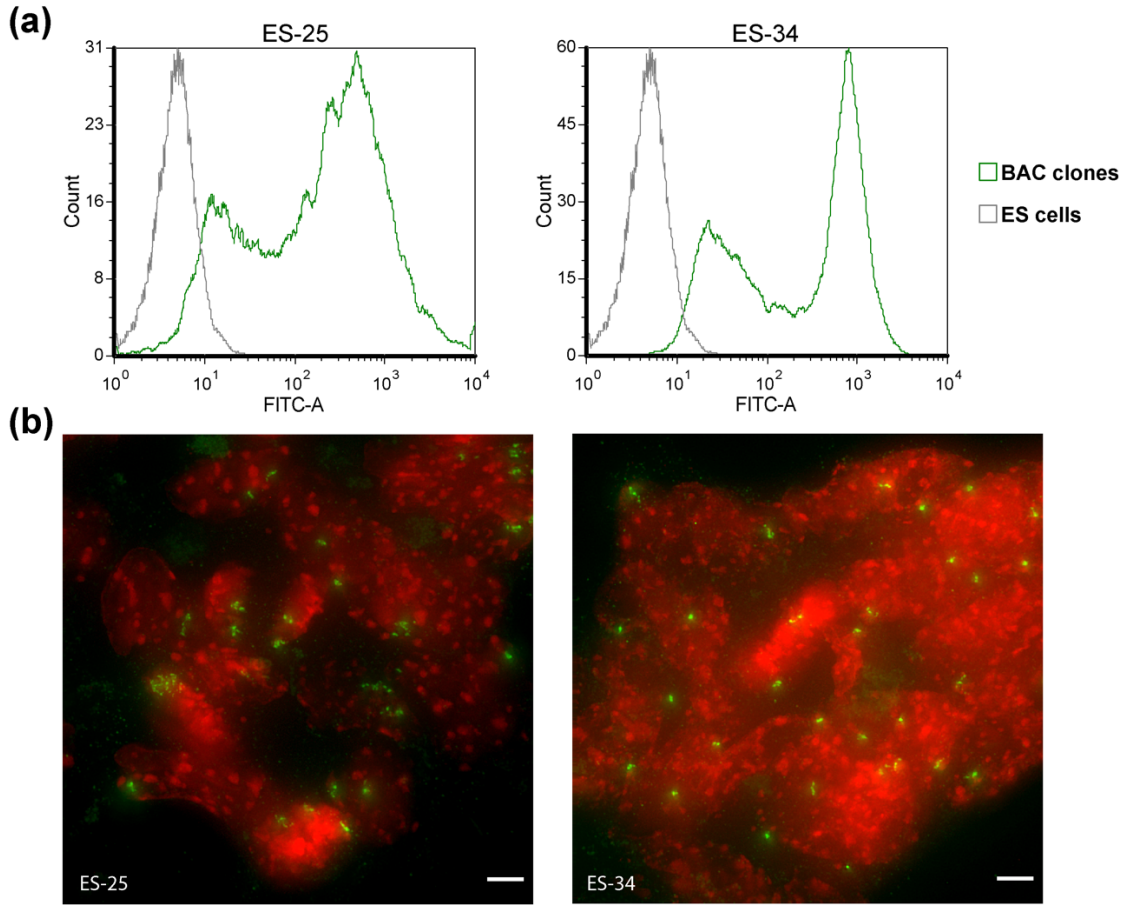

Figure B.1. GAPDH BAC did not form episomes in mouse ES cells. (a) Broad distribution of GFP histograms of two mouse ES cell clones containing GAPDH BAC (green). The GAPDH BAC and the clones were described in a previous study (Chaturvedi *et al.*, 2018). Gray: untransfected mouse ES cells. (b) z-projections of DNA FISH images of the two clones in panel a. GAPDH BAC probes were used for DNA FISH. Red- DNA DAPI staining; Green- FISH signal; Scale bars = 5 μm.

REFERENCES

Chaturvedi, P. *et al.* (2018) ‘Stable and reproducible transgene expression independent of proliferative or differentiated state using BAC TG-EMBED.’, *Gene therapy*, 25(5), pp. 376–391.

APPENDIX C: VISUALIZING LARGE GENOMIC REGIONS WITH CRISPR/CAS9 SYSTEM

INTRODUCTION

In order to study chromatin large-scale structure and spatial organization, specific genomic regions need to be visualized. Ideally, the visualization method should work in both fixed and living cells, so as to study the dynamics of the chromatin during various physiological activities. The ideal visualization method should also be able to label any genomic region of any size, without disturbing chromatin structure of the labeled region.

Robust DNA Fluorescence *in situ* Hybridization (FISH) protocols have been developed to label both small genomic loci ~100s kb in size, and large genomic regions several Mbs in size or even the whole chromatin (Cremer *et al.*, 2008; Boyle *et al.*, 2011). However, DNA FISH has several disadvantages: the procedure is long, spanning several days; cells must be fixed; chromatin structure is not well preserved after DNA FISH procedure, due to multiple hash treatments to ensure successfully hybridization of FISH probes to the genomic DNA, including HCl extraction, freeze-thaw cycles and heat denaturation.

Alternative visualization methods using a fluorescent repressor-operator system (Belmont and Straight, 1998; Tasan *et al.*, 2018) can visualize genomic loci both in fixed and living cells with minimal disruption of the chromatin. However, such systems only mark the genomic insertion sites of the tandem operator repeats, and do not light up long chromatin trajectories. Moreover, it is time consuming to establish cell clones with

operator repeats stably integrated at the desired genomic loci while expressing an optimal level of fluorescent repressors.

The revolutionary genome editing tool, type II clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated caspase 9 (Cas9) system derived from *Streptococcus pyogenes* (Doudna and Charpentier, 2014; Hsu, Lander and Zhang, 2014) has also been used for visualization of genomic loci, where a fluorescently tagged nuclease-deficient Cas9 (dCas9) is used instead of Cas9. While most studies visualize genomic loci by establishing cell lines simultaneously expressing fluorescently tagged dCas9 and single-guide RNAs (sgRNAs) (Chen *et al.*, 2013; Tanenbaum *et al.*, 2014; B. Chen *et al.*, 2018; Gu *et al.*, 2018; Ma *et al.*, 2018), one study showed efficient labeling of genomic loci by incubating methanol-acidic acid fixed cells with in-vitro assembled dCas9/sgRNA complexes (Deng *et al.*, 2015). This so-called CASFISH method is rapid and robust, and thus has the potential to substitute traditional DNA FISH if it has similar or even better chromatin preservation comparing to DNA FISH.

Currently, highly repetitive sequences consisting of hundreds of repeats have been visualized efficiently both by expressing dCas9/sgRNA and by CASFISH. However, visualizing non-repetitive regions with CRISPR/CAS systems remains challenging (Chen *et al.*, 2013; Deng *et al.*, 2015), as it requires a large number of different sgRNAs, depending on the fluorescent tag of the system and the length of the genomic region to be visualized, to achieve a sufficient signal-to-noise ratio for microscopy detection. Current strategies to circumvent the need of large quantities of different sgRNAs include appending RNA-aptamer motifs to sgRNAs (Qin *et al.*, 2017; Ma *et al.*, 2018), and

inserting tandem repeats of CRISPR targetable DNA sequences into the genomic loci to be visualized (Gu *et al.*, 2018; Y. Chen *et al.*, 2018).

Here I show successful visualization of a large BAC transgene array ~20 Mb in size with pooled 13 sgRNAs using the CASFISH protocol. Each of these sgRNAs is specific to a naturally occurring, low-copy number tandem repetitive sequence unique to the BAC constituting the transgene array. The total copy number of these tandem repeats is ~100 copies per BAC. While the endogenous loci corresponding to the region contained within the BAC could not be visualized by these sgRNAs, this method could be applied to certain large genomic regions with sufficient number of unique tandem repeats.

RESULTS

Efficiency of CASFISH

As CASFISH (Deng *et al.*, 2015) is a newly developed method, I first tested the published protocol with three repetitive sequences, mouse major satellite repeats (MSR), human *MUC4* gene exon 2 (E2) and intron 1 (I1) repeats, which were used as demonstrations in the paper (Deng *et al.*, 2015).

The standard CASFISH protocol from the paper (Deng *et al.*, 2015) was used with BSA removed from the blocking/reaction buffer (Deng *et al.*, 2015) and with no blocking step. As expected, MSR CASFISH generated strong signals over the chromocenters of the nuclei of mouse NIH 3T3 cells (Figure C.1a). Both E2 and I1 CASFISH generated puncta in the majority of nuclei (Figure C.1b-c). The E2 puncta had

good signal-to-noise ratio, while the I1 puncta were weaker than E2 and resembled background puncta (Figure C.1b-c). Increasing concentration of dCas9/sgRNA in the incubation step could not improve signal-to-noise ratio, indicating limited binding of dCas9/sgRNA to the genomic DNA (data not shown).

Although it is claimed that the E2 repeats have ~400 copies of repeats, while the I1 repeats ~90 copies (Chen *et al.*, 2013), CasFinder (Aach, Mali and Church, 2014), a software for identifying specific Cas9 targets in genomes, identified 125 (97% of all binding sites over the whole hg19 genome) and 127 (99% of all binding sites over the whole hg19 genome) binding sites over *MUC4* gene for E2 and I1 sgRNA, respectively. However, there is a significant difference in the spacing of the E2 and I1 sgRNA binding sites over the *MUC4* gene. The median distances of adjacent binding sites (5' to 5') were 48 bp and 15 bp for E2 and I1, respectively. Thus it is likely that both the number of sgRNA binding sites and the spacing of the sgRNA binding sites contributes to the efficiency of CASFISH.

CASFISH with formaldehyde fixation was tested using MSR sgRNA, as formaldehyde fixation could potentially preserve the chromatin better than methanol/acetic acid fixation. A low concentration of formaldehyde and short incubation time, and a 15 min HCl treatment were required to detect the signal (Figure C.2).

CASFISH of DHFR BAC transgene array with 12 pooled sgRNAs

Previous studies have successfully visualized a 5-kb nonrepetitive region of *MUC4* gene both by CASFISH using 73 sgRNAs (Deng *et al.*, 2015) and by expressing dCas9 and 26~73 sgRNAs in the cells (Chen *et al.*, 2013). For visualizing this 5 kb

nonrepetitive region, CASFISH requires an additional stringent wash step. Here I show an alternative way to visualize large non-repetitive genomic regions by CASFISH, by designing sgRNAs targeting naturally occurring low-copy tandem repetitive sequences scattered in non-repetitive genomic regions.

While the number of high copy number tandem repetitive sequences is limited in the genome, low copy tandem repeats are abundant. 12 tandem repeats over a BAC containing the mouse *Dhfr-Msh3* locus (DHFR BAC) were identified using Tandem Repeat Finder (Benson, 1999). 13 sgRNAs targeting these tandem repeats were then designed using CasFinder (Aach, Mali and Church, 2014). These sgRNAs have a total of 100 binding sites over the DHFR BAC and 104 over the whole mm9 genome.

The sgRNAs were tested on an NIH 3T3 clone containing a large lac operator repeats (LacO) tagged DHFR BAC transgene array (~20 Mb) and expressing EGFP-dimer lac repressor-NLS (nuclear localization signal) fusion protein (EGFP-LacI). As expected, CASFISH using the pooled sgRNA targeting the tandem repeats successfully visualized the trajectory of the DHFR BAC transgene array (Figure C.3).

MATERIALS AND METHODS

Cell culture and DHFR BAC

Mouse NIH 3T3 cells and human U2OS cells were cultured with Dulbecco's modified Eagle medium (DMEM, with 4.5 g/L D-glucose, 4 mM L-glutamine, 1 mM sodium pyruvate and 3.7 g/L $NaHCO_3$) supplemented with 10% HyClone Bovine Growth Serum (GE Healthcare Life Sciences, Cat. # SH30541.03).

The DHFR-c27 BAC (Bian and Belmont, 2010) containing a 256-mer lac operator direct repeat (LacO) and a CMV-mRFP-SV40-ZeoR expression cassette was derived from the CITB-057L22 BAC (DHFR BAC) containing mouse chr13:92,992,156-93,161,185 (mm9). NIH 3T3 cell clone DHFR-c27-13 FULGW containing ~117 copies of the DHFR-c27 BAC and expressing an EGFP-dimer lac repressor-NLS (nuclear localization signal) fusion protein (EGFP-LacI) from lentivirus FULGW was established in previous studies (Bian and Belmont, 2010; Sinclair *et al.*, 2010). The cell clone was maintained with 75 μg/ml Zeocin (Thermo Fisher Scientific).

dCas9 constructs and purification

Purified dCas9-tdTomato and dCas9-mNeonGreen (Lane *et al.*, 2015) were a gift from Dr. Rebecca Heald (University of California at Berkeley, Berkeley, CA, USA).

Plasmid pBZ03-NLS-dCas9-Halo-NLS-6xHis was derived from plasmid pET302-6His-dCas9-Halo, a gift from Timothée Lionnet (Addgene plasmid # 72269 ; http://n2t.net/addgene:72269 ; RRID:Addgene_72269) using NEBuilder HiFi DNA Assembly (New England Biolabs). All oligos used for construction were listed in Table C.1.

To construct the intermediate plasmid pBZ02-NLS-dCas9-Halo, the following fragments were used: a 1.5 kb PCR products amplified from plasmid pET302-6His-dCas9-Halo using primer pair BZ#327/328, a connecting oligo BZ#335, and a 9248 bp vector backbone derived from digesting plasmid pET302-6His-dCas9-Halo with NcoI-HF and PmeI (New England Biolabs). To construct plasmid pBZ03-NLS-dCas9-Halo-NLS-6xHis, the following fragments were used: a 177 bp PCR products amplified from

plasmid pET302-6His-dCas9-Halo using primer pair BZ#329/330, a connecting oligo BZ#336 and a 10621 bp vector backbone derived from digesting plasmid pBZ02-NLS-dCas9-Halo with XmaI and XhoI (New England Biolabs).

NLS-dCas9-Halo-NLS-6xHis protein was expressed in Rosetta (DE3) (MilliporeSigma), grown in LB medium at ~18°C for ~18 h following induction with 0.2 mM IPTG. Cells were lysed in Buffer A (20 mM Tris-Cl, pH=8.0; 500 mM NaCl) by sonication. Clarified lysate was applied to a HisTrap HP column (GE Healthcare). The bound protein was washed by increasing concentration of imidazole up to 50 mM and eluted by increasing imidazole up to 100 mM. The eluted protein was concentrated using an Ultra-15 100K centrifugal filter (Amicon), exchanged into Buffer C (20mM Tris-Cl pH=8.0; 200 mM KCl; 10 mM $MgCl_2$; 10% glycerol) using a PD-10 column (GE Healthcare) x2, and concentrated again using an Ultra-4 3K centrifugal filter (Amicon). Protein concentration was measured using A280 (molecular weight = 195.5 kDa, extinction coefficient = 169600 $M^{-1}cm^{-1}$).

NLS-dCas9-Halo-NLS-6xHis protein was fluorescently labeled by incubating 10-40 μM protein with Janelia Fluor 549-Halo tag ligands (a gift from Janelia Research Campus) or with Alexa Fluor 660-Halo tag ligands (Promega) in Buffer C at 4°C for ~16 h, with protein/dye molar ratio = 1:4 – 1:8. Free dyes were removed by 50K centrifugal filter (Amicon).

Synthesis of MSR, E1 and E2 sgRNA templates

The targeting sequences of MSR, E1 and E2 sgRNA and the sgRNA scaffold sequence were from previous studies (Chen *et al.*, 2013; Deng *et al.*, 2015). Each sgRNA

template was synthesized by PCR using two overlapping oligos synthesized by Integrated DNA Technologies (Table C.2). One oligo contains a T7 promoter, a sgRNA targeting sequence and a partial sgRNA scaffold (MSR-sgRNA-forward, MUC4-E2-sgRNA-forward, or MUC4-I1-sgRNA-forward). The other oligo contains a partial sgRNA scaffold (Linker3-bottom). PCR used the following recipe: 1x Q5 reaction buffer, 200 μM dNTPs (New England Biolabs), 0.5 μM each oligo, and 0.02 U/μl Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs), and the following running protocol: 98°C 30 s; 35 cycles of 98°C 10 s, 71°C 30 s, and 72°C 5 s; 72°C 2 min; and 4°C HOLD.

Design of DHFR BAC sgRNA targeting sequences

Tandem repeats over the DHFR BAC were searched by Tandem Repeat Finder (Benson, 1999) (version 4.09, Sequence: mm9_dna range=chr13:92992156-93161185 5'pad=0 3'pad=0 strand=+ repeatMasking=none; Parameters: 2 7 7 80 10 50 500). Repeats with consensus size > 20 and period size ≥ 30 were merged into repetitive regions. 12 repetitive regions with length ≥ 120 bp were selected and named DHFRt1, DHFRt2, …, DHFRt12 (Table C.3). The summed length of the 12 repetitive regions was ~5 kb.

20 nt and 23 nt candidate sgRNA targeting sequences over DHFRt1, DHFRt2, …, DHFRt12 were searched by CasFinder (Aach, Mali and Church, 2014) (distribution: 05122014, options: -D -A 200 -R -g mm9). Total numbers of binding sites of each candidate targeting sequences over the DHFR BAC regions and over the whole mm9 genome, and the spacing of binding sites over the DHFR BAC regions were calculated

from the output files of CasFinder. 13 candidate targeting sequences were manually selected according to the following conditions: 5' end starting with "G"; maximum binding sites over the DHFR BAC; minimum binding sites over the rest of the genome; maximum spacing. DHFRt3 has two candidate targeting sequences and each of the rest repetitive regions has one. Next, to create the final oligos BZ-lib-01, -02, …, -26, PAM sequences were removed from the 13 candidate targeting sequences; a single "G" was added to candidate targeting sequences that did not begin with "GG"; a reverse complimentary sequence of each candidate targeting sequence was created; and finally overhangs compatible to the BsaI sites of plasmid pBZ01 were added to each sequence. Sequences of BZ-lib-01, -02, …, -26 were listed in Table C.4.

Construction of DHFR BAC sgRNA template plasmids

The strategy for constructing sgRNA template plasmids was similar to previous studies (Hwang *et al.*, 2013; Zhou *et al.*, 2017). A 147 bp sequence containing a T7 promoter positioned upstream of an optimized sgRNA scaffold sequence (Chen *et al.*, 2013) was synthesized and cloned into plasmid pIDTSMART-KAN by Integrated DNA Technologies, yielding plasmid pBZ01 (Figure C.4 and Data C.1). Two BsaI sites were positioned in between the T7 promoter and the sgRNA scaffold sequence, where sgRNA targeting sequences could be cloned into to make a complete sgRNA template. A BstBI and a HindIII sites were positioned immediately downstream of sgRNA scaffold sequence for linearization of the plasmid for in vitro transcription.

Two oligos were designed for each sgRNA targeting sequence so that the annealed oligos contain overhangs compatible to the BsaI sites of plasmid pBZ01. The

oligos (BZ-lib-01, -02, …, -26) were synthesized by Integrated DNA Technologies, phosphorylated using T4 Polynucleotide Kinase (New England Biolabs) and annealed in a thermocycler using the following protocol: 98°C 2 min; 20°C 2 min (Ramp = 0.02°/s); 4°C pause. Plasmid pBZ01 was digested by BsaI-HF (New England Biolabs), gel purified using QIAquick Gel Extraction Kit (Qiagen) and ligated with the annealed oligos using Quick Ligation Kit (New England Biolabs), yielding plasmid pBZ01-1/2, pBZ01-3/4, …, pBZ01-25/26.

Synthesis of sgRNAs

sgRNAs were synthesized by in vitro transcription using MEGAscript T7 Transcription Kit (Thermo Fisher Scientific) according to manufacturer's instructions, with ~150 ng PCR products or ~450 ng linearized plasmids per 10 μl reaction as templates and ~16 h incubation time. The sgRNAs were purified using MEGAclear Transcription Clean-up Kit (Thermo Fisher Scientific) according to manufacturer's instructions. The yield was ~50 μg purified sgRNA per 10 μl reaction.

CASFISH

CASFISH used the standard protocol (Deng *et al.*, 2015) with small modifications. Cells grown on coverslips were fixed with methanol/acetic acid (1v:1v) at -20°C for ~20 min and washed with DPBS (Chapter 2) x3. Fluorescently labeled dCas9 (1-25 nM) and sgRNA (dCas9:sgRNA = 1:2 to 1:4 molar ratio) were incubated in B/R Buffer (20 mM HEPES, pH 7.5; 100 mM KCl; 5 mM $MgCl_2$; 5% glycerol; 0.1% TWEEN and 1 mM DTT) for ~10 min at room temperature (RT) and then on ice until the

next step. The fixed cells were incubated with the dCas9/sgRNA mixture for 30 min at 37°C, washed with B/R Buffer for 5 min x3, and mounted.

Fixation with formaldehyde was tested using MSR sgRNA. Cells were fixed with 0.5% paraformaldehyde/DPBS for 5 min at RT, permeabilized with 0.5% Triton X-100/DPBS and incubated in 0.1 N HCl for 15 min. Rest steps were the same as above.

Code Availability

All computational scripts used for designing sgRNA targeting sequences are available at https://bitbucket.org/Binhui/sgrna-design/src.

FIGURES AND TABLES

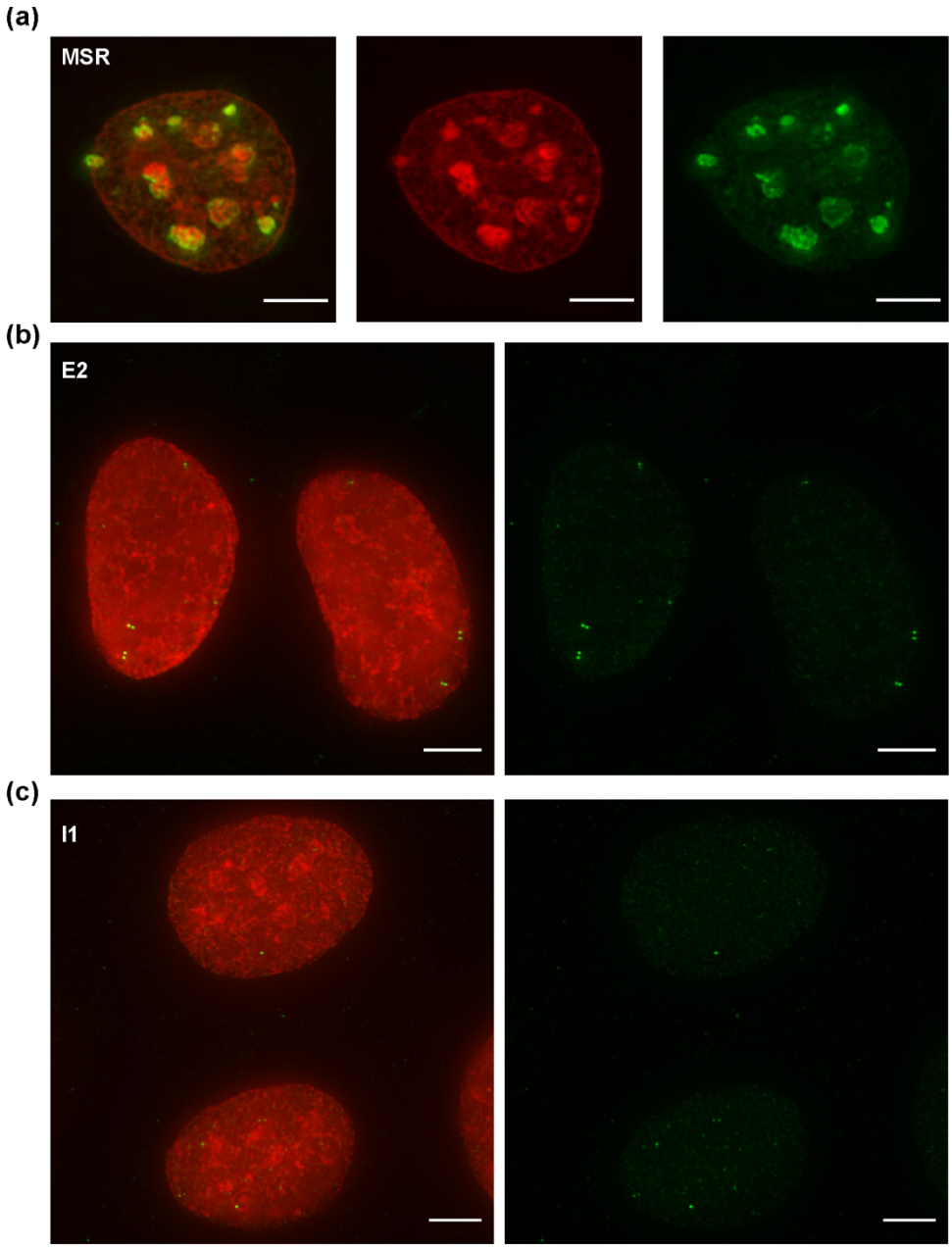

Figure C.1. CASFISH of MSR, E2 and I1. (a) CASFISH using MSR sgRNA over NIH 3T3 cells. Gamma = 0.5 was applied to green channels. (b-c): CASFISH using E2 (b) or I1 (c) sgRNA over U2OS cells. Red- DNA DAPI staining; Green- CASFISH signal; Scale bars = 5 μm.

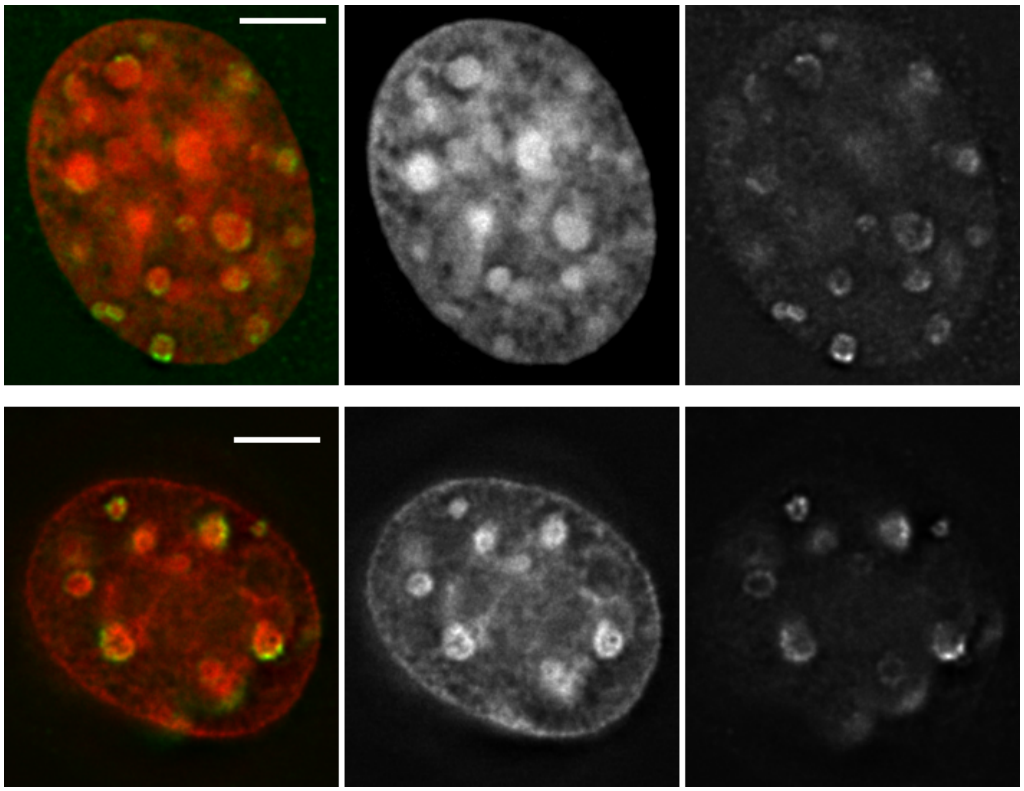

Figure C.2. CASFISH of MSR using formaldehyde fixation (top panel) and methanol/acetic acids fixation (bottom panel). NIH 3T3 cells were used. Bottom panel reuses the image for Figure C.1. Red and Middle panel- DNA DAPI staining; Green and right panel- CASFISH signal; Scale bars = 4 μm.

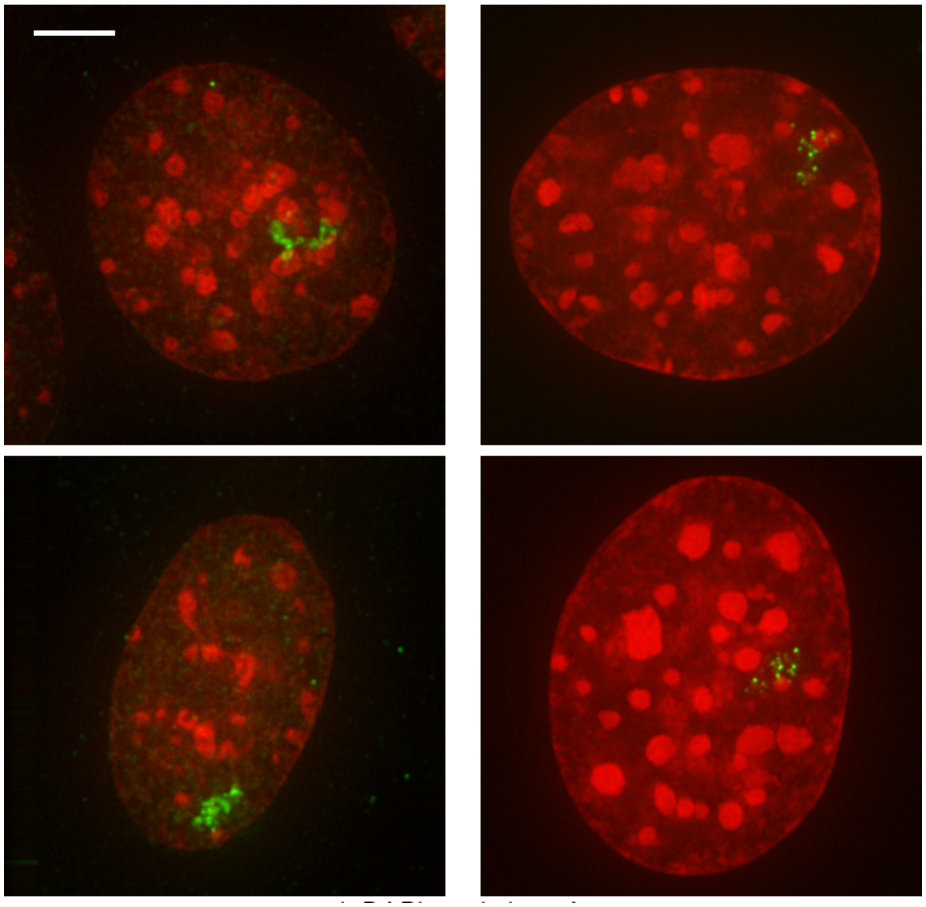

Figure C.3. DHFR BAC transgene array in a NIH 3T3 clone DHFR-c27-13 visualized by CASFISH (left) vs by EGFP-LacI (right). Maximum intensity projections are shown. Red- DAPI; Green- CASFISH or EGFP-LacI; Scale bar = 4 μm.

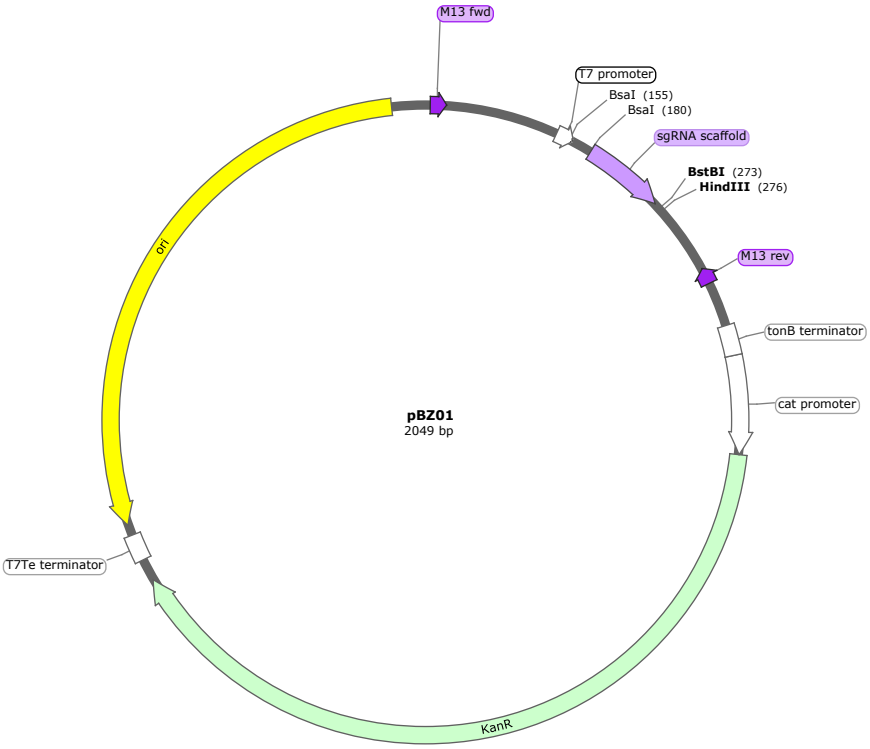

Figure C.4. Map of plasmid pBZ01.

Table C.1. Oligo sequences for constructing pBZ03-NLS-dCas9-Halo-NLS-6xHis.

Oligo name	Oligo sequence (5'-3')
BZ#327	cccaagaagaagaggaagGTGGATAAGAAATACTCAATAGG
BZ#328	CACCTTTATCGACAACTTC
BZ#329	CTGCCTAACTGCAAGGCTG
BZ#330	atgatgcgagccaccgcccactttgcgtttcttttcggcagacgggtgccggagccaccgCCGGAAATCTCCAGAGTAGAC
BZ#335	ACAATTCCCCTCTAGAAATAATTTTGTTTAAACTTTAAGAAGGAGATATACTATG cccaagaagaagaggaaggtgGATAAGAA
BZ#336	aaacgcaaagtgggcggtggctcgcatcatcatcatcatcactaaTAGCTCGAGATCGATGATATTCGAGCCTAGGTATAAT

Table C.2. Oligo sequences for synthesizing MSR, E2 and I1 sgRNA templates.

Oligo name	**Oligo sequence (5'-3')**
MSR-sgRNA-forward	gaaatTAATACGACTCACTATAGGCCATATTCCACGTCCTACAGGTTTAAGAGCTATGCTGG AAACAGCATAGC
MUC4-E2-sgRNA-forward	gaaatTAATACGACTCACTATAGGAAGTGTCGGTGACAGGAAGAGTTTAAGAGCTATGCTGG AAACAGCATAGC
MUC4-I1-sgRNA-forward	gaaatTAATACGACTCACTATAGGAAGGTATGGGTGTGGAAGGTATGTTTAAGAGCTATGCTGG AAACAGCATAGC
Linker3-bottom	AAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTAAACTTG CTATGCTGTTTCCAGCATAGCTCTTAAAC

Table C.3. Tandem repetitive regions over the DHFR BAC.

Repeat name	Repeat length	Repeat mm9 coordinate
DHFRt1	869bp	chr13:92999581-93000449
DHFRt2	156bp	chr13:93018743-93018898
DHFRt3	404bp	chr13:93028764-93029167
DHFRt4	150bp	chr13:93029976-93030125
DHFRt5	193bp	chr13:93070622-93070814
DHFRt6	343bp	chr13:93081847-93082189
DHFRt7	1378bp	chr13:93085813-93087190
DHFRt8	298bp	chr13:93097220-93097517
DHFRt9	275bp	chr13:93106886-93107160
DHFRt10	165bp	chr13:93124956-93125120
DHFRt11	494bp	chr13:93134619-93135112
DHFRt12	335bp	chr13:93139363-93139697

Table C.4. Oligo sequences used to synthesize the targeting sites of DHFR sgRNA templates.

Oligo name	Oligo sequence (5'-3')	Tandem Repeat
BZ-lib-01	taGGGTCACGGAGGCTGGGTGTGTG	DHFRt1
BZ-lib-02	aaacCACACACCCAGCCTCCGTGAC	
BZ-lib-03	taGGATGCGCGGCGGGCCTTGGTGG	DHFRt10
BZ-lib-04	aaacCCACCAAGGCCCGCCGCGCAT	
BZ-lib-05	tagGTTTAATGGCAGAACCCACACAG	DHFRt11
BZ-lib-06	aaacCTGTGTGGGTTCTGCCATTAAA	
BZ-lib-07	tagGTATTCATTTAGTGCACTCCCAT	DHFRt12
BZ-lib-08	aaacATGGGAGTGCACTAAATGAATA	
BZ-lib-09	tagGAGAGCAGTCTATGCATGAC	DHFRt2
BZ-lib-10	aaacGTCATGCATAGACTGCTCT	
BZ-lib-11	taGGCCTCACAATGAATGTCCAGTA	DHFRt3
BZ-lib-12	aaacTACTGGACATTCATTGTGAGG	
BZ-lib-13	taGGACCTCTCAATGTCCAATA	DHFRt3
BZ-lib-14	aaacTATTGGACATTGAGAGGT	
BZ-lib-15	taGGATGACAGATAGATCGCTG	DHFRt4
BZ-lib-16	aaacCAGCGATCTATCTGTCAT	
BZ-lib-17	taGGCCTTAAGCGTGCTCACAGATA	DHFRt5
BZ-lib-18	aaacTATCTGTGAGCACGCTTAAGG	
BZ-lib-19	tagGAAGGAGGAGGAGGCTTTACAGC	DHFRt6
BZ-lib-20	aaacGCTGTAAAGCCTCCTCCTCCTT	
BZ-lib-21	taGGTCAGCTGTCCACACCTCACAG	DHFRt7
BZ-lib-22	aaacCTGTGAGGTGTGGACAGCTGA	
BZ-lib-23	tagGTAGCTTGTCTGTTGCAGTGTGT	DHFRt8
BZ-lib-24	aaacACACACTGCAACAGACAAGCTA	
BZ-lib-25	tagGAGGTCATGTGAGTGTGAGTAGG	DHFRt9
BZ-lib-26	aaacCCTACTCACACTCACATGACCT	

Data C.1. Plasmid pBZ01 sequence. Sequence and annotations of plasmid pBZ01 saved as a GenBank file (Data_C.1.gb).

REFERENCES

Aach, J., Mali, P. and Church, G. M. (2014) 'CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes', *bioRxiv*, pp. 1–8.

Belmont, A. S. and Straight, A. F. (1998) 'In vivo visualization of chromosomes using lac operator-repressor binding.', *Trends in cell biology*, 8(3), pp. 121–4.

Benson, G. (1999) 'Tandem repeats finder: a program to analyze DNA sequences.', *Nucleic acids research*, 27(2), pp. 573–80.

Bian, Q. and Belmont, A. S. (2010) 'BAC TG-EMBED: one-step method for high-level, copy-number-dependent, position-independent transgene expression.', *Nucleic acids research*, 38(11), p. e127.

Boyle, S. *et al.* (2011) 'Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis.', *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 19(7), pp. 901–9.

Chen, B. *et al.* (2013) 'Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system', *Cell*. Elsevier Inc., 155(7), pp. 1479–1491.

Chen, B. *et al.* (2018) 'Efficient labeling and imaging of protein-coding genes in living cells using CRISPR-Tag.', *Nature communications*, 9(1), p. 5065.

Chen, Y. *et al.* (2018) 'Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler', *Journal of Cell Biology*, 217(11), pp. 4025–4048.

Cremer, M. *et al.* (2008) 'Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes.', *Methods in molecular biology (Clifton, N.J.)*, 463, pp. 205–39.

Deng, W. *et al.* (2015) 'CASFISH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells', *Proceedings of the National Academy of Sciences*, p. 201515692.

Doudna, J. A. and Charpentier, E. (2014) 'The new frontier of genome engineering with CRISPR-Cas9', *Science*, 346(6213), pp. 1258096–1258096.

Gu, B. *et al.* (2018) 'Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements', *Science*, 359(6379), pp. 1050–1055.

Hsu, P. D., Lander, E. S. and Zhang, F. (2014) 'Development and Applications of CRISPR-Cas9 for Genome Engineering', *Cell*, 157(6), pp. 1262–1278.

Hwang, W. Y. *et al.* (2013) 'Efficient genome editing in zebrafish using a CRISPR-Cas system', *Nat Biotechnol*. Nature Publishing Group, 31(3), pp. 227–229.

Lane, A. B. B. *et al.* (2015) 'Enzymatically Generated CRISPR Libraries for Genome Labeling and Screening', *Developmental Cell*. Elsevier Inc., 34(3), pp. 373–378.

Ma, H. *et al.* (2018) 'CRISPR-Sirius: RNA scaffolds for signal amplification in genome imaging', *Nature Methods*, 15(11), pp. 928–931.

Qin, P. *et al.* (2017) 'Live cell imaging of low- and non-repetitive chromosome loci using CRISPR-Cas9', *Nature Communications*. The Author(s), 8, p. 14725.

Sinclair, P. *et al.* (2010) 'Dynamic plasticity of large-scale chromatin structure revealed by self-assembly of engineered chromosome regions.', *The Journal of cell biology*, 190(5), pp. 761–76.

Tanenbaum, M. E. *et al.* (2014) 'A Protein-Tagging System for Signal Amplification in Gene Expression and Fluorescence Imaging', *Cell*. Elsevier Inc., 159(3), pp. 635–646.

Tasan, I. *et al.* (2018) 'CRISPR/Cas9-mediated knock-in of an optimized TetO repeat for live cell imaging of endogenous loci.', *Nucleic acids research*, 46(17), p. e100.

Zhou, Y. *et al.* (2017) 'Painting a specific chromosome with CRISPR/Cas9 for live-cell imaging.', *Cell research*, 27(2), pp. 298–301.

APPENDIX D: DESIGN OF OLIGO LIBRARIES FOR VISUALIZING LARGE GENOMIC REGIONS WITH CRISPR/CAS9 SYSTEM

To visualize large non-repetitive mammalian genomic regions, pooled sgRNAs tiling the genomic regions need to be synthesized. Here I represent a strategy for synthesizing such sgRNA libraries. First, a pool of oligos containing the sgRNA targeting sequences is synthesized. Next, the oligos were PCR amplified and cloned into plasmid pBZ01 (Appendix C), which contains a T7 promoter upstream of an sgRNA scaffold, with two BsaI sites positioned in between the T7 promoter and the sgRNA scaffold for cloning. Finally, the plasmid libraries are used for synthesis of sgRNA libraries by in vitro transcription. This strategy is similar to published studies (Bao *et al.*, 2018), except here multiple libraries are synthesized within one oligo pool.

A pool of 12461 oligos was designed for visualizing 30 different mouse and human genomic regions (Table D.1). Each oligo has the following structures: FFFFFFFFFFFFFFFFFFggtctcataggNNNNNNNNNNNNNNNNNNgtttagagaccRRRRRRRRRRRRRRRRRRR.

The FFFFFFFFFFFFFFFFFF and RRRRRRRRRRRRRRRRRR represent 18 nt PCR primer binding sites for amplifying each sgRNA targeting sequence library. The ggNNNNNNNNNNNNNNNNNN represents 20 nt sgRNA targeting sequence. The small letters are BsaI sites.

sgRNAs candidate targeting sites were searched by CasFinder (Aach, Mali and Church, 2014). The candidate targeting sites were then selected with the following steps: 1) targeting sites starting with "GN" or "NG" were selected; 2) targeting sites containing

BsaI sites (GGTCTC|GAGACC) were removed; 3) targeting sites with 35%~ 65% GC% were selected; 4) targeting sites containing 'TTTT..' were removed; 5) targeting sites were selected so that distance to adjacent targeting sites (5'-5') >= 46.

This targeting sites selection did not check HindIII or BstBI sites, which are used to linearize the final plasmid library for in vitro transcription. The percentage of bad oligos that will be cut by HindIII or BstBI is listed in Table D.2.

Primers for amplifying each library are designed by the following steps: 1) 18mers were derived from a library of orthogonal 25mers (Xu *et al.*, 2009) with the following conditions: GC content in between 40% and 60%, not containing BsaI sites, T_m ≥57°C and ≤ 63°C, hairpin T_m ≤ 0°C, homodimer T_m ≤ 0°C, homodimer dG >= -3500, no XXXX…, no XYXYXY…; 2) mis-priming of the 18mers against the sgRNA targeting sequences were checked by primer3-2.3.7 (Untergasser *et al.*, 2012), oligos with mis-priming scores ≤ 10 were selected; 3) mis-priming of the 18mers against the sgRNA targeting sequences were checked again by blast 2.6.0; 4) mis-priming of the 18mers among themselves were checked by blast 2.6.0; 5) oligos were filtered by Sequence Manipulation Suite: PCR Primer Stats (http://www.bioinformatics.org/sms2/pcr_primer_stats.html); 6) oligos with minimum heterodimer formation and similar Tm (calculated by NEB Tm Calculator (https://tmcalculator.neb.com/#!/main) using Q5 polymerase) were paired.

Detailed information of each oligo is in Data D.1. Detailed information of the primers for library amplification is in Data D.2. Computational scripts for designing the oligo sequences are available at https://bitbucket.org/Binhui/sgrna-design/src.

Table D.1. Information of the libraries in the oligo pool.

	Library name	Number of oligos	Targeting sequence	Description	Specificity
1	Chr12_A2_peak-h-lib_1	34	RP11-154B14 (BAC)	A2 peak of TSA seq,	human
2	Chr18_A2_pInv-h-lib_1	97	RP11-650D13 (BAC)	A2 peak in valley of TSA seq	human
3	Chr18_A2_pInv-h-lib_2	97			
4	Chr2_A1_pTos-h-lib_1	2520	chr2:22925891-27325691 (hg19)	periphery to speckle trajectory	human
5	Chr2_A1_sTop-h-lib_1	4398	chr2:27325671-34336642 (hg19)	speckle to periphery trajectory	human
6	Chr2_A2_peak-h-lib_1	119	RP11-598F14 (BAC)	A2 peak of TSA seq	human
7	Chr2_A2_v1-h-lib_1	1749	chr2:36000000- 40950032 (hg19)	A2 valley of TSA seq	human
8	Chr2_A2_v2-h-lib_1	1130			
9	Chr3_A2_peak-h-lib_1	156	RP11-922N10 (BAC)	A2 peak of TSA seq	human
10	Chr4_R_lib1	239	chr4:144965603- 147315930 (mm9)	mouse naturally repetitive region	mouse
11	Chr6_A2_pInv-h-lib_1	90	RP11-20H19 (BAC)	A2 peak in valley of TSA seq	human
12	Chr6_A2_pInv-h-lib_2	90			
13	COL1A1_BAC-h-lib_1	120	RP11-267M22 (BAC)	COL1A1 BAC	human
14	COL1A1_BAC-h&m-lib_1	95			human and mouse
15	COL1A1_BAC-h&m-lib_2	94			
16	DHFR_BAC-m-lib_1	116	CITB-057L22 (BAC)	DHFR BAC	mouse
17	DHFR_BAC-m-lib_2	116			
18	GAPDH_BAC-h-lib_1	125	RP11-369N23 (BAC)	GAPDH BAC	human
19	GAPDH_BAC-h&m-lib_1	126			human and mouse
20	GAPDH_BAC-h&m-lib_2	125			
21	HBB_BAC-h-lib_1	76	CTD-2643I7 (BAC)	HBB BAC	human
22	HBB_BAC-h&m-lib_1	59			human and mouse
23	HBB_BAC-m-lib_1	103			mouse
24	HBB_BAC-m-lib_2	102			
25	HBB_BAC-m-lib-3	102			
26	HSPH1_BAC-h-lib_1	130	RP11-173P16 (BAC)	Heat shock BAC	human
27	HSPH1_BAC-h-lib_2	130			
28	K13_BAC-h-lib_1	11	CTD-2207K13 (BAC)	2207K13 BAC	human
29	K13_BAC-h&m-lib_1	17			human and mouse
30	K13_BAC-m-lib_1	95			mouse

Table D.2. Percentage of oligos with HindIII or BstBI sites.

Library name	Total number of oligos	Oligos with HindIII sites	Oligos with BstBI sites	Oligos with HindIII or BstBI sites	Bad oligos%
Chr12_A2_peak-h-lib-1	34	0	0	0	0.00%
Chr18_A2_pInv-h-lib-1	97	0	0	0	0.00%
Chr18_A2_pInv-h-lib-2	97	1	0	1	1.03%
Chr2_A1_pTos-h-lib-1	2520	15	3	18	0.71%
Chr2_A1_sTop-h-lib-1	4398	20	10	30	0.68%
Chr2_A2_peak-h-lib-1	119	1	0	1	0.84%
Chr2_A2_v1-h-lib-1	1749	6	3	9	0.51%
Chr2_A2_v2-h-lib-1	1130	4	4	8	0.71%
Chr3_A2_peak-h-lib-1	156	0	0	0	0.00%
Chr4_R_lib1	239	0	0	0	0.00%
Chr6_A2_pInv-h-lib-1	90	0	0	0	0.00%
Chr6_A2_pInv-h-lib-2	90	0	0	0	0.00%
COL1A1_BAC-h-lib-1	120	0	1	1	0.83%
COL1A1_BAC-h&m-lib-1	95	0	1	1	1.05%
COL1A1_BAC-h&m-lib-2	94	0	0	0	0.00%
DHFR_BAC-m-lib-1	116	0	0	0	0.00%
DHFR_BAC-m-lib-2	116	2	0	2	1.72%
GAPDH_BAC-h-lib-1	125	1	0	1	0.80%
GAPDH_BAC-h&m-lib-1	126	0	0	0	0.00%
GAPDH_BAC-h&m-lib-2	125	2	1	3	2.40%
HBB_BAC-h-lib-1	76	0	0	0	0.00%
HBB_BAC-h&m-lib-1	59	1	0	1	1.69%
HBB_BAC-m-lib-1	103	1	0	1	0.97%
HBB_BAC-m-lib-2	102	1	0	1	0.98%
HBB_BAC-m-lib-3	102	2	0	2	1.96%
HSPH1_BAC-h-lib-1	130	2	0	2	1.54%
HSPH1_BAC-h-lib-2	130	2	0	2	1.54%
K13_BAC-h-lib-1	11	0	0	0	0.00%
K13_BAC-h&m-lib-1	17	0	0	0	0.00%
K13_BAC-m-lib-1	95	1	1	2	2.11%

Data D.1. Information of each oligo. A tab delimited text file (Data_D.1.txt) providing the following information: library name, sequence name, oligo sequence without primer binding sites, amplification primer ID, forward primer name, alternative forward primer name, forward primer sequence, reverse primer name, alternative reverse primer name, reverse primer sequence, reverse complementary sequence of reverse primer, and complete oligo sequence of each oligo.

Data D.2. Primers for amplifying the libraries in the oligo pool. A tab delimited text file (Data_D.2.txt) providing the following information: amplification primer ID, forward primer name, alternative forward primer name, forward primer sequence, reverse primer name, alternative reverse primer name, reverse primer sequence, and library name for each oligo library.

REFERENCES

Aach, J., Mali, P. and Church, G. M. (2014) 'CasFinder: Flexible algorithm for identifying specific Cas9 targets in genomes', *bioRxiv*, pp. 1–8.

Bao, Z. *et al.* (2018) 'Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision', *Nature Biotechnology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 36, p. 505.

Untergasser, A. *et al.* (2012) 'Primer3—new capabilities and interfaces', *Nucleic Acids Research*, 40(15), pp. e115–e115.

Xu, Q. *et al.* (2009) 'Design of 240,000 orthogonal 25mer DNA barcode probes.', *Proceedings of the National Academy of Sciences of the United States of America*, 106(7), pp. 2289–94.